H. Kawai et al.
10/16/03
Q77945
1 of 1

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００２年１０月１７日

出願番号
Application Number: 特願２００２－３０２５８５

[ST. 10/C]: [ＪＰ２００２－３０２５８５]

出願人
Applicant(s): 日本電気株式会社

２００３年　８月２０日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０６８０１２

【書類名】　特許願
【整理番号】　35001170
【提出日】　平成14年10月17日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　河合　英紀
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　福島　俊一
【特許出願人】
　【識別番号】　000004237
　【氏名又は名称】　日本電気株式会社
【代理人】
　【識別番号】　100088959
　【弁理士】
　【氏名又は名称】　境　廣巳
【手数料の表示】
　【予納台帳番号】　009715
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9002136
【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　ハイパーテキスト検査装置および方法並びにプログラム

【特許請求の範囲】

【請求項１】

ハイパーテキストデータベースを対象とし、論理的なリンク不整合の箇所を検出することを特徴としたハイパーテキスト検査装置。

【請求項２】

リンク元表記とリンク先ページ内容との間に不整合が生じている箇所、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所、同一ページ内および周辺ページ内の複数のリンク元表記の間にスタイルの不統一が生じている箇所、リンク元表記のないリンク箇所、ループを形成するリンクの系列であって且つ該リンクの系列に対応するリンク元表記がすべて同一トピックに関わる箇所、のうちの少なくとも１種類の論理的なリンク不整合の箇所を検出することを特徴とした請求項１に記載のハイパーテキスト検査装置。

【請求項３】

ハイパーテキストを構成するページおよびリンクに関する情報を格納する情報記憶手段と、前記情報記憶手段を参照して論理的なリンク不整合の箇所を検出する条件判定手段とを備えることを特徴としたハイパーテキスト検査装置。

【請求項４】

ハイパーテキストを構成するページおよびリンクに関する情報を収集する情報収集手段と、前記ページおよびリンクに関する情報を格納する情報記憶手段と、前記情報記憶手段を参照して論理的なリンク不整合の箇所を検出する条件判定手段とを備えることを特徴としたハイパーテキスト検査装置。

【請求項５】

前記条件判定手段によって検出された箇所に関する訂正候補を計算する候補計算手段を備えることを特徴とした請求項３又は４に記載のハイパーテキスト検査装置。

【請求項６】

前記条件判定手段によって検出された箇所の重要度を出力する重要度計算手段を備えることを特徴とした請求項５に記載のハイパーテキスト検査装置。

【請求項７】

前記条件判定手段が検出したリンク不整合の箇所と前記候補計算手段が計算した訂正候補とに基づいて前記ハイパーテキストを更新する訂正反映手段を備えることを特徴とした請求項５または６に記載のハイパーテキスト検査装置。

【請求項８】

前記重要度計算手段で計算された重要度、前記条件判定手段によって検出された箇所の数、総リンク数に対する前記条件判定手段によって検出された箇所の数の割合、のうちの１ファクタもしくは複数のファクタの組み合わせによって前記ハイパーテキストに関するトータルスコアを計算して出力するトータルスコア計算手段を備えることを特徴とした請求項６に記載のハイパーテキスト検査装置。

【請求項９】

前記条件判定手段によって検出された箇所の重要度を出力する重要度計算手段を備えることを特徴とした請求項３または４に記載のハイパーテキスト検査装置。

【請求項１０】

前記重要度計算手段で計算された重要度、前記条件判定手段によって検出された箇所の数、総リンク数に対する前記条件判定手段によって検出された箇所の数の割合、のうちの１ファクタもしくは複数のファクタの組み合わせによって前記ハイパーテキストに関するトータルスコアを計算して出力するトータルスコア計算手段とを備えることを特徴とした請求項９に記載のハイパーテキスト検査装置。

【請求項１１】

前記条件判定手段は、リンクに関する情報を特定の条件でグループ化し、グループから外れたリンクに関する情報をリンク不整合の箇所として検出するようにした、請求項３、４、９または１０に記載のハイパーテキスト検査装置。

【請求項１２】

前記条件判定手段は、リンク元表記とリンク先ページ内容との間に不整合が生じている箇所を検出するようにした、請求項３、４、９または１０に記載のハイパ

ーテキスト検査装置。

【請求項１３】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第２の不適正スコア、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第３の不適正スコア、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算される第４の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにした、請求項３、４、９または１０に記載のハイパーテキスト検査装置。

【請求項１４】

前記条件判定手段は、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するようにした、請求項３、４、９または１０に記載のハイパーテキスト検査装置。

【請求項１５】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク先ページ内容に含まれる移動や期限切れの告知表現を検出することで計算される第２の不適正スコア、または、（３）リンク元表記あるいはリンク先ページ内容に含まれる有効期間の記述と現在日時を比較することで計算される第３の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにした、請求項３、４、９または１０に記載のハイパーテキスト検査装置。

【請求項１６】

前記条件判定手段は、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所を検出するようにした、請求項３、４、９または１０に記載のハイパーテキスト検査装置。

【請求項１７】

前記条件判定手段は、同一ページ内または周辺ページ内の複数のリンク元表記の間にスタイルの不統一が生じている箇所を検出するようにした、請求項３、４、９または１０に記載のハイパーテキスト検査装置。

【請求項１８】

前記条件判定手段は、リンクに関する情報を特定の条件でグループ化し、グループから外れた特異なリンクに関する情報をリンク不整合の箇所として検出するようにした、請求項５乃至８の何れか１項に記載のハイパーテキスト検査装置。

【請求項１９】

前記候補計算手段は、前記特異なリンクに関する情報を、不整合でない他のリンクと同一にするような訂正候補を求めるようにした、請求項１８に記載のハイパーテキスト検査装置。

【請求項２０】

前記条件判定手段は、リンク元表記とリンク先ページ内容との間に不整合が生じている箇所を検出するようにした、請求項５乃至８の何れか１項に記載のハイパーテキスト検査装置。

【請求項２１】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第２の不適正スコア、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第３の不適正スコア、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算される第４の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにし、前記候補計算手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算されるリンク元表記の訂正候補、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算されるリンク先の訂正候補、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される

リンク先の訂正候補、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算されるリンク元表記の訂正候補、のうちのいずれか１種類もしくは複数種類を訂正候補とするようにした、請求項５乃至８の何れか１項に記載のハイパーテキスト検査装置。

【請求項２２】

前記条件判定手段は、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するようにした、請求項５乃至８の何れか１項に記載のハイパーテキスト検査装置。

【請求項２３】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク先ページ内容に含まれる移動や期限切れの告知表現を検出することで計算される第２の不適正スコア、または、（３）リンク元表記あるいはリンク先ページ内容に含まれる有効期間の記述と現在日時を比較することで計算される第３の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにし、前記候補計算手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算されるリンク元表記の訂正候補、または、（２）リンク先ページ内容から移動先に関する情報を抽出することで計算されるリンク先の訂正候補、のうちの一方もしくは両方を訂正候補とするようにした、請求項５乃至８の何れか１項に記載のハイパーテキスト検査装置。

【請求項２４】

前記条件判定手段は、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所を検出するようにし、前記候補計算手段は、検出された箇所とリンク先ページが同一の複数リンクについてリンク元表記を比較することでリンク元表記の訂正候補を計算するようにした、請求項５乃至８の何れか１項に記載のハイパーテキスト検査装置。

【請求項２５】

前記条件判定手段は、同一ページ内および周辺ページ内の複数のリンク元表記の

間にスタイルの不統一が生じている箇所を検出するようにし、前記候補計算手段は、検出された箇所を含むページおよび周辺ページ内の複数のリンク元表記のスタイルを比較することでリンク元表記のスタイルの訂正候補を計算するようにした、請求項５乃至８の何れか１項に記載のハイパーテキスト検査装置。

【請求項２６】

前記情報収集手段は、ハイパーテキストを構成するページおよびリンクに関する情報の収集を繰り返し実行し、前記情報格納手段は、複数の時刻における前記ページおよびリンクに関する情報を格納するようにした、請求項４乃至１０のいずれか１項に記載のハイパーテキスト検査装置。

【請求項２７】

前記条件判定手段は、前記情報格納手段を参照し、内容に変更が生じたページに対する被リンク数の時間変化やリンク元表記の種類の時間変化を計算することで、リンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するようにした、請求項２６に記載のハイパーテキスト検査装置。

【請求項２８】

前記条件判定手段は、リンク元表記のないリンクを検出するようにした、請求項３乃至１０のいずれか１項に記載のハイパーテキスト検査装置。

【請求項２９】

前記条件判定手段は、リンク元表記として文字列も画像も設定されていないリンク、あるいは、リンク元表記として目に付きにくい色やサイズを用いた文字列や画像が設定されたリンクを検出するようにした、請求項３乃至１０のいずれか１項に記載のハイパーテキスト検査装置。

【請求項３０】

前記条件判定手段は、ループを形成するリンクの系列であって、かつ、該リンクの系列に対応するリンク元表記がすべて同一トピックに関わるものを検出するようにした、請求項３乃至１０のいずれか１項に記載のハイパーテキスト検査装置。

【請求項３１】

前記重要度計算手段は、（１）検出された箇所の誤り／不適切の種類、（２）検

出された箇所の誤り／不適切の確度、（３）検出された箇所を含むページの被リンク数、（４）検出された箇所を含むページに対するユーザからのアクセス実績、（５）検出された箇所を含むページのハイパーテキストにおける階層レベル、のうちの１ファクタもしくは複数のファクタの組み合わせによって重要度を計算するようにした、請求項６または９に記載のハイパーテキスト検査装置。

【請求項３２】

前記重要度計算手段は、検出された箇所の重要度を計算し、該重要度に基づいて、前記検出された箇所の出力件数や出力方法を制御するようにした、請求項６、９または３１に記載のハイパーテキスト検査装置。

【請求項３３】

前記情報収集手段は、リンク元表記が画像の場合に文字認識を行うことによって該リンク元表記に対応する文字列も抽出し、前記ページおよびリンクに関する情報の１つとして前記情報格納手段に登録するようにした、請求項４乃至３２のいずれか１項に記載のハイパーテキスト検査装置。

【請求項３４】

Webサイトに存在するハイパーテキストを検査対象とした、請求項１乃至３３のいずれか１項に記載のハイパーテキスト検査装置。

【請求項３５】

ハイパーテキストデータベースを対象に、リンク元表記やリンク関係における誤り箇所あるいは不適切箇所を検出する条件を指定させるステップと、該条件に合致する箇所について（１）リンク元表記、（２）リンク元ページの識別情報、（３）リンク先ページの識別情報、という３項目を一覧表示するステップとを含むことを特徴としたハイパーテキスト検査方法。

【請求項３６】

前記一覧表示するステップにおいて、条件に合致した箇所を、（１）リンク元表記、（２）リンク元ページの識別情報、（３）リンク先ページの識別情報、という３項目のいずれかをソートキーとして一覧表示するようにした、請求項３５に記載のハイパーテキスト検査方法。

【請求項３７】

ハイパーテキストデータベースを対象に、リンク元表記やリンク関係における誤り箇所あるいは不適切箇所を検出する条件を指定させるステップと、該条件に合致する箇所について（１）リンク元表記、（２）リンク元ページの識別情報、（３）リンク先ページの識別情報、という３項目をスクリーンに一覧表示するステップと、前記スクリーン上で（１）または（２）または（３）の項目を訂正させるステップと、スクリーン上で訂正された結果にしたがって前記ハイパーテキストデータベースを一括修正するステップとを含むことを特徴としたハイパーテキスト検査方法。

【請求項３８】

検査対象とするハイパーテキストデータベースを指定するステップを含むことを特徴とした請求項３５乃至３７のいずれか１項に記載のハイパーテキスト検査方法。

【請求項３９】

Webサイトを構成するページおよびリンクに関する情報を収集する情報収集ステップと、前記情報収集ステップの結果を参照して論理的なリンク不整合の箇所を検出する条件判定ステップと、前記条件判定ステップによって検出された箇所の重要度およびWebサイトのトータルスコアを計算する重要度計算ステップとを、指定されたWebサイトを対象として定期的に自動実行し、前記指定されたWebサイトに関する前記トータルスコアの時間変化を通知することを特徴としたハイパーテキスト検査方法。

【請求項４０】

Webサイトを構成するページおよびリンクに関する情報を収集する情報収集ステップと、前記情報収集ステップの結果を参照して論理的なリンク不整合の箇所を検出する条件判定ステップと、前記条件判定ステップによって検出された箇所の重要度およびWebサイトのトータルスコアを計算する重要度計算ステップとを、指定されたWebサイトを対象として定期的に自動実行し、前記指定されたWebサイトに関する前記トータルスコアあるいは前記検出された箇所の重要度があらかじめ定めた条件を満たした場合にアラートを通知することを特徴としたハイパーテキスト検査方法。

【請求項４１】

Webサイトを構成するページおよびリンクに関する情報を収集する情報収集ステップと、前記情報収集ステップの結果を参照して論理的なリンク不整合の箇所を検出する条件判定ステップと、前記条件判定ステップによって検出された箇所の重要度およびWebサイトのトータルスコアを計算する重要度計算ステップとを、指定された複数のWebサイトを対象として自動実行し、前記指定された複数のWebサイトの間の前記トータルスコアの順位付け結果を通知することを特徴としたハイパーテキスト検査方法。

【請求項４２】

ハイパーテキストを構成するページおよびリンクに関する情報を格納する情報記憶手段を備えたコンピュータを、前記情報記憶手段を参照して論理的なリンク不整合の箇所を検出する条件判定手段、として機能させるハイパーテキスト検査プログラム。

【請求項４３】

情報記憶手段を備えたコンピュータを、ハイパーテキストを構成するページおよびリンクに関する情報を収集して前記情報記憶手段に格納する情報収集手段、前記情報記憶手段を参照して論理的なリンク不整合の箇所を検出する条件判定手段、として機能させるハイパーテキスト検査プログラム。

【請求項４４】

前記コンピュータを、さらに、前記条件判定手段によって検出された箇所に関する訂正候補を計算する候補計算手段、として機能させる請求項４３に記載のハイパーテキスト検査プログラム。

【請求項４５】

前記コンピュータを、さらに、前記条件判定手段によって検出された箇所の重要度を出力する重要度計算手段、として機能させる請求項４４に記載のハイパーテキスト検査プログラム。

【請求項４６】

前記コンピュータを、さらに、前記条件判定手段が検出したリンク不整合の箇所と前記候補計算手段が計算した訂正候補とに基づいて前記ハイパーテキストを更

新する訂正反映手段、として機能させる請求項４４または４５に記載のハイパーテキスト検査プログラム。

【請求項４７】

前記コンピュータを、さらに、前記重要度計算手段で計算された重要度、前記条件判定手段によって検出された箇所の数、総リンク数に対する前記条件判定手段によって検出された箇所の数の割合、のうちの１ファクタもしくは複数のファクタの組み合わせによって前記ハイパーテキストに関するトータルスコアを計算して出力するトータルスコア計算手段、として機能させる請求項４５に記載のハイパーテキスト検査プログラム。

【請求項４８】

前記コンピュータを、さらに、前記条件判定手段によって検出された箇所の重要度を出力する重要度計算手段、として機能させることを特徴とした請求項４２または４３に記載のハイパーテキスト検査プログラム。

【請求項４９】

前記コンピュータを、さらに、前記重要度計算手段で計算された重要度、前記条件判定手段によって検出された箇所の数、総リンク数に対する前記条件判定手段によって検出された箇所の数の割合、のうちの１ファクタもしくは複数のファクタの組み合わせによって前記ハイパーテキストに関するトータルスコアを計算して出力するトータルスコア計算手段、として機能させる請求項４８に記載のハイパーテキスト検査プログラム。

【請求項５０】

前記条件判定手段は、リンクに関する情報を特定の条件でグループ化し、グループから外れたリンクに関する情報をリンク不整合の箇所として検出するようにした、請求項４２、４３、４８または４９に記載のハイパーテキスト検査プログラム。

【請求項５１】

前記条件判定手段は、リンク元表記とリンク先ページ内容との間に不整合が生じている箇所を検出するようにした、請求項４２、４３、４８または４９に記載のハイパーテキスト検査プログラム。

【請求項５２】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第２の不適正スコア、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第３の不適正スコア、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算される第４の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにした、請求項４２、４３、４８または４９に記載のハイパーテキスト検査プログラム。

【請求項５３】

前記条件判定手段は、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するようにした、請求項４２、４３、４８または４９に記載のハイパーテキスト検査プログラム。

【請求項５４】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク先ページ内容に含まれる移動や期限切れの告知表現を検出することで計算される第２の不適正スコア、または、（３）リンク元表記あるいはリンク先ページ内容に含まれる有効期間の記述と現在日時を比較することで計算される第３の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにした、請求項４２、４３、４８または４９に記載のハイパーテキスト検査プログラム。

【請求項５５】

前記条件判定手段は、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所を検出するようにした、請求項４２、４３、４８または４９に記載のハイパーテキスト検査プログラム。

【請求項５６】

前記条件判定手段は、同一ページ内または周辺ページ内の複数のリンク元表記の間にスタイルの不統一が生じている箇所を検出するようにした、請求項４２、４３、４８または４９に記載のハイパーテキスト検査プログラム。

【請求項５７】

前記条件判定手段は、リンクに関する情報を特定の条件でグループ化し、グループから外れた特異なリンクに関する情報をリンク不整合の箇所として検出するようにした、請求項４４乃至４７の何れか１項に記載のハイパーテキスト検査プログラム。

【請求項５８】

前記候補計算手段は、前記特異なリンクに関する情報を、不整合でない他のリンクと同一にするような訂正候補を求めるようにした、請求項５７に記載のハイパーテキスト検査プログラム。

【請求項５９】

前記条件判定手段は、リンク元表記とリンク先ページ内容との間に不整合が生じている箇所を検出するようにした、請求項４４乃至４７の何れか１項に記載のハイパーテキスト検査プログラム。

【請求項６０】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第２の不適正スコア、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第３の不適正スコア、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算される第４の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにし、前記候補計算手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算されるリンク元表記の訂正候補、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算されるリンク先の訂正候補、または、（３）リンク元ページおよびリンク

元表記が同一の複数リンクについてリンク先ページを比較することで計算されるリンク先の訂正候補、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算されるリンク元表記の訂正候補、のうちのいずれか１種類もしくは複数種類を訂正候補とするようにした、請求項４４乃至４７の何れか１項に記載のハイパーテキスト検査プログラム。

【請求項６１】

前記条件判定手段は、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するようにした、請求項４４乃至４７の何れか１項に記載のハイパーテキスト検査プログラム。

【請求項６２】

前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク先ページ内容に含まれる移動や期限切れの告知表現を検出することで計算される第２の不適正スコア、または、（３）リンク元表記あるいはリンク先ページ内容に含まれる有効期間の記述と現在日時を比較することで計算される第３の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにし、前記候補計算手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算されるリンク元表記の訂正候補、または、（２）リンク先ページ内容から移動先に関する情報を抽出することで計算されるリンク先の訂正候補、のうちの一方もしくは両方を訂正候補とするようにした、請求項４４乃至４７の何れか１項に記載のハイパーテキスト検査プログラム。

【請求項６３】

前記条件判定手段は、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所を検出するようにし、前記候補計算手段は、検出された箇所とリンク先ページが同一の複数リンクについてリンク元表記を比較することでリンク元表記の訂正候補を計算するようにした、請求項４４乃至４７の何れか１項に記載のハイパーテキスト検査プログラム。

【請求項６４】

前記条件判定手段は、同一ページ内および周辺ページ内の複数のリンク元表記の間にスタイルの不統一が生じている箇所を検出するようにし、前記候補計算手段は、検出された箇所を含むページおよび周辺ページ内の複数のリンク元表記のスタイルを比較することでリンク元表記のスタイルの訂正候補を計算するようにした、請求項４４乃至４７の何れか１項に記載のハイパーテキスト検査プログラム。

【請求項６５】

前記情報収集手段は、ハイパーテキストを構成するページおよびリンクに関する情報の収集を繰り返し実行し、前記情報格納手段は、複数の時刻における前記ページおよびリンクに関する情報を格納するようにした、請求項４３乃至４９のいずれか１項に記載のハイパーテキスト検査プログラム。

【請求項６６】

前記条件判定手段は、前記情報格納手段を参照し、内容に変更が生じたページに対する被リンク数の時間変化やリンク元表記の種類の時間変化を計算することで、リンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するようにした、請求項６５に記載のハイパーテキスト検査プログラム。

【請求項６７】

前記条件判定手段は、リンク元表記のないリンクを検出するようにした、請求項４２乃至４９のいずれか１項に記載のハイパーテキスト検査プログラム。

【請求項６８】

前記条件判定手段は、リンク元表記として文字列も画像も設定されていないリンク、あるいは、リンク元表記として目に付きにくい色やサイズを用いた文字列や画像が設定されたリンクを検出するようにした、請求項４２乃至４９のいずれか１項に記載のハイパーテキスト検査プログラム。

【請求項６９】

前記条件判定手段は、ループを形成するリンクの系列であって、かつ、該リンクの系列に対応するリンク元表記がすべて同一トピックに関わるものを検出するようにした、請求項４２乃至４９のいずれか１項に記載のハイパーテキスト検査プログラム。

【請求項７０】

前記重要度計算手段は、（１）検出された箇所の誤り／不適切の種類、（２）検出された箇所の誤り／不適切の確度、（３）検出された箇所を含むページの被リンク数、（４）検出された箇所を含むページに対するユーザからのアクセス実績、（５）検出された箇所を含むページのハイパーテキストにおける階層レベル、のうちの１ファクタもしくは複数のファクタの組み合わせによって重要度を計算するようにした、請求項４５または４８に記載のハイパーテキスト検査プログラム。

【請求項７１】

前記重要度計算手段は、検出された箇所の重要度を計算し、該重要度に基づいて、前記検出された箇所の出力件数や出力方法を制御するようにした、請求項４５、４８または７０に記載のハイパーテキスト検査プログラム。

【請求項７２】

前記情報収集手段は、リンク元表記が画像の場合に文字認識を行うことによって該リンク元表記に対応する文字列も抽出し、前記ページおよびリンクに関する情報の１つとして前記情報格納手段に登録するようにした、請求項４３乃至７１のいずれか１項に記載のハイパーテキスト検査プログラム。

【請求項７３】

Webサイトに存在するハイパーテキストを検査対象とした、請求項４２乃至７２のいずれか１項に記載のハイパーテキスト検査プログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明はハイパーテキスト検査装置に関し、特にリンク元表記やリンク関係における誤り箇所を検知するハイパーテキスト検査装置に関する。

【０００２】

【従来の技術】

近年、企業、団体および個人がインターネットのサイト上に電子化した情報を公開することが多くなった。これらサイト上に公開される情報の多くはハイパーテ

キストである。

【０００３】

ハイパーテキストとは、ハイパーリンク（リンク）で構造化された文書集合のことであり、文書をノードとし、文書間にリンクをはった構造を持つ。ハイパーテキストの代表例が、WWW(World Wide Web)である。WWWは図2(a)の文書101ようにHTML(Hyper Text Markup Language)形式で記述されたハイパーテキストの集合であり、リンク及びアンカー文字列は<A>タグによってマークされる。図2(a)の文書101では、<A>タグのhref属性に文書102、103、104の識別情報が指定されている。文書の識別情報は、WWWでは通常URLまたはwebアドレスと呼ばれるが、本発明では、単にアドレスと呼ぶこととする。また、<A>タグで挟まれた、「GX0011」「GX0012」「GX0013」などの文字列は一般にアンカー文字列と呼ばれる。ただし、<A>タグで画像ファイルを挟むこともあるため、本発明では<A>タグで挟まれた文字列または画像をリンク元表記と呼んで同様に扱うことにする。

【０００４】

文書101で記述されている<A>タグの属性には、href属性の他にもtarget属性、style属性なども存在する。target属性は、リンク先の文書を表示するウィンドウを指定するための属性である。また、style属性はリンクのリンク元表記を表示する際のフォントの大きさや色、強調表現などを指定するための属性である。図2(a)の文書101をブラウザで閲覧すると図2(b)の文書101のように表示される。図2(b)の文書101では、リンク元表記「GX0011」をクリックすることにより、リンク201を介して文書102にアクセスすることができる。同様にリンク元表記「GX0012」をクリックするとリンク202を介して文書103に、リンク元表記「GX0013」をクリックするとリンク203を介して文書104にアクセスすることができる。

【０００５】

なお、ハイパーテキストの代表例としてWWWについて説明したが、本発明は対象をWWWに限定したものではない。ハイパーテキストはHTMLだけではなく、XML(Extensible Markup Language)、SGML(Standard Generalized Markup Language)等を用いて記述することも可能である。

【０００６】

また、本発明では「利用者」という用語の混乱をさけるために企業・団体・個人のサイトを訪れてハイパーテキストを閲覧する人を「訪問者」、本発明を利用してハイパーテキストを管理する人を「管理者」と呼ぶことにする。

【０００７】

インターネットに公開される情報量の増大とともにハイパーテキストの管理は複雑困難になっており、リンク元表記が不適切なリンクや、リンク先を誤ったリンクなど、リンク不整合の件数が増大している。リンク不整合には、おおきく物理的不整合と論理的不整合の2種類に分類できる。

【０００８】

物理的不整合は、リンク先のテキストがない、リンク先のサーバーがダウンしている、など、物理的にリンク先にアクセス不可能な不整合である。これら物理的不整合では、文書にアクセスした時点でサーバーやクライアントがエラーを返す。

【０００９】

論理的不整合は、間違った製品情報へのリンクや、期限切れのキャンペーンへのリンクなど、物理的にはアクセス可能であっても、論理的な誤りを生じている不整合である。これら論理的不整合では、リンク先にテキストは存在しており、リンク先のサーバーにも異常はないため、文書にアクセスした時点でエラーは発生しない。しかし、間違いリンクによって訪問者は混乱してしまったり、期限切れのキャンペーンへ応募する訪問者が発生して管理者が対応に苦慮するなど、その影響は物理的不整合に劣らず大きい。論理的不整合の例としては、(1)リンクの張り間違い、(2)期限切れ情報へのリンク、(3)リンク元表記の不統一、(4)リンク元表記のスタイルの不統一、(5)幽霊リンク、(6)ループリンクなどが挙げられる。以下に、論理的不整合の各例について図面を参照して詳細に説明する。

【００１０】

(1)リンクの張り間違い

リンクの張り間違いは図3に示すように、リンク元表記で期待される内容と、リンク先のテキストの内容がずれている場合の不整合である。図3では、リンク211、212、213、214のリンク元表記はすべて「GX0011」で同じである。また、文書1

11、112、113のリンク211、212、213のリンク先はいずれも文書116である。そのため文書111、112、113を閲覧した訪問者は、リンク元表記「GX0011」で期待される通り、GX0011の紹介情報である文書116にアクセスすることができる。ところが、リンク214のリンク先は間違ってGX0012の製品紹介である文書117を指定している。そのため、文書114を閲覧した訪問者は、リンク元表記「GX0011」で期待される情報とは別の製品紹介を見せられることになり、混乱してしまう。

【００１１】

また、リンク211、212、213、215のリンク先はすべて文書116である。ところが、リンク215のリンク元表記は間違って「GX0012」と記述されている。そのため、文書115を閲覧した訪問者は、リンク元表記「GX0012」で期待される情報とは別の製品紹介を見せられることになり、混乱してしまう。

【００１２】

また、文書115から張られている2つのリンク215、216はどちらもリンク元表記が「GX0012」となっている。ところが、それぞれのリンク先は文書116、117と異なっているため、文書115を閲覧した訪問者は同じリンク元表記にもかかわらず異なる文書をたどることになり、混乱してしまう。

【００１３】

なお、ここではリンクの張り間違いの例として製品情報へのリンク間違いを説明したが、他にも、英語版文書と日本語版文書間でのリンクの張り間違いや、全く関係のないページへの間違いリンクなどの不整合もリンクの張り間違いに含む。

【００１４】

(2)期限切れ情報へのリンク

期限切れ情報へのリンクは図4に示すように、期限切れのキャンペーンや閉鎖したサービスへのリンクが残っている場合の不整合である。図4(a)では、文書125で2002年7月20日～2002年8月31日までの期間限定でキャンペーンが行われている。また、文書121、122、123、124からはキャンペーンページである文書125へ、同じリンク元表記「入会無料」でそれぞれリンク221、222、223、224が張られている。一方、図4(b)では、キャンペーン期間を終了したために文書125では終了を告知している。また、文書121、122、123ではキャンペーンページである文書1

25へのリンクを削除している。にもかかわらず文書124では、リンクを削除し忘れているためにリンク元表記「入会無料」で文書125へのリンク224が残っている。そのため、文書124を閲覧した訪問者は、リンク元表記で「入会無料」を期待しても、そのサービスは受けられないことになってしまう。

【００１５】

なお、ここでは期限切れ情報へのリンクの例として期限切れキャンペーンへのリンクを説明したが、他にも、最初にリンクしてあった文書のアドレスが移転し、元のアドレスに別の内容の文書が置かれることにより発生する不整合も期限切れ情報へのリンクに含む。また、最初から期間を限定していなくても、何らかの理由によりリンク先のサービスが終了されたり、サイトが閉鎖されたりすることによって生じる不整合も期限切れ情報へのリンクに含む。ただし、文書が期限切れになって削除されている場合は、アクセス時にエラーが発生するため、物理的不整合に含む。また、期限切れリンクは間違いリンクの一種で考えてもよいが、本発明では、間違いリンクの中でも特にリンク先が期限切れとなったものを、期限切れリンクとして区別している。

【００１６】

(3)リンク元表記の不統一

リンク元表記の不統一は図5に示すように、リンク元表記が統一されず揺らぎがある場合の不整合である。図5では、文書131、132、133、134から、文書135へのリンクが張られている。リンク231、232、233のリンク元表記はいずれも「GX Series」である。ところが、リンク234のリンク元表記が「gX Series」となっている。そのため、文書134を閲覧した訪問者は、リンク元表記で「GX Series」とは異なる「gX Series」が存在するのかと勘違いしてリンク234をたどってしまう。

【００１７】

なお、ここではリンク元表記の不統一の例としてリンク元表記の大文字・小文字の揺らぎを説明したが、他にも、「トップページ」「Topページ」などの英語・カタカナの表記の揺らぎ、「ヴァイオリン」「バイオリン」などのカタカナ表記の揺らぎ、「スイカ」「すいか」などのカタカナ・ひらがなの表記の揺らぎ、「イベント情報」「セミナー情報」などのあいまいな類似表現による表記の揺らぎ

、「Series」「Selies」などのスペルミスなどもリンク元表記の不統一に含む。

【００１８】

(4)リンク元表記のスタイルの不統一

リンク元表記のスタイルの不統一は図6に示すように、リンクのstyle属性やtarget属性が異なっているために、リンクの見え方が異なったり、リンクをクリックした時の効果が異なる場合の不整合である。図6(a)では、文書141で4つのリンクが指定されており、そのうち3件ではリンクをクリックした時にポップアップウィンドウにリンク先のページを表示するよう、target属性「_blank」が指定されている。そのため、図6(b)のように文書141をブラウザで閲覧している訪問者は、リンク241、242、243のリンク先文書を、文書141を開いたまま次々とポップアップウィンドウで閲覧できる。特にリンク集など、リンク先を閲覧し、戻ってまた別のリンク先を閲覧することが多い文書では、このようにポップアップウィンドウにリンク先のページを表示すると便利なことが多い。一方、リンク244にはtarget属性が指定されていないため、リンクをクリックした時に文書が切り替わる設定になっている。そのため、訪問者がリンク244をクリックすると文書が切り替わってしまい、文書141へ戻るためのリンクを探したり、ブラウザの戻るボタンを使わなければならなくなってしまう。

【００１９】

なお、ここではリンク元表記のスタイルの不統一の例としてtarget属性の不統一を説明したが、他にも、style属性が不統一のために、一部のリンクの色が異なっていたり、一部のリンクの強調表現の有無が不統一だったりする不整合もリンク元表記のスタイルの不統一に含む。

【００２０】

(5)幽霊リンク

幽霊リンクは図7に示すように、文書のHTML表記ではリンクが指定されているにも関わらず、ブラウザで閲覧するとそのリンクの存在に気がつかない場合の不整合である。図7(a)では、見出しを表す「GXシリーズ在庫状況」という文字列と、テーブルを表す<TABLE>タグの間に、<A>タグがあり、リンク先にHIDDEN_URLを指定している。ところが、リンク元表記として<A>タグの間に文字列や画像などを

何も挟んでいないために、ブラウザで閲覧した時に図7(b)のように見出しと表の間にリンクがあることがわからない。このようなリンクは、クローラーでたどることは容易だが、管理者によるチェックは困難である。仮に、HIDDEN_URLが顧客リストなど機密ファイルを指していると、クローラーで容易に機密情報が取得できてしまう一方で、人間ではその漏洩に気が付かないといった問題が起こるおそれもある。

【００２１】

なお、ここでは幽霊リンクの例としてリンク元表記が何も指定されない場合を説明したが、他にも、リンク元表記に透明な画像が指定されていたり、非常に小さな画像や文字が指定されていたり、背景と同じ色の画像や文字が指定されるなど、ブラウザでの目視確認が困難な場合の不整合も幽霊リンクに含む。また、見えてはいても、リンク元表記のリンクスタイルが本文と同じ色で何も強調表現がなく、リンクと本文との見分けがつかない場合も、ブラウザでの目視確認が困難であるため幽霊リンクに含む。

【００２２】

(6)ループリンク

ループリンクは図8に示すように、ある情報を求めてリンクをたどっていくと、元のページに戻ってしまう場合の不整合である。図8では、文書161から文書162へ、リンク元表記「プレゼントのお知らせ」でリンク261が張られている。また、文書162から文書163へ、リンク元表記「デジカメプレゼント」でリンク262が張られている。さらに、文書163から文書161へ、リンク元表記「プレゼントはこちら」でリンク263が張られている。例えば文書161を閲覧した訪問者が「プレゼントのお知らせ」に興味を持ってリンク261をたどったとする。すると、文書162でもリンク元表記が「デジカメプレゼント」であるリンク262があるため、その先にプレゼントの詳細情報があると期待して文書163にアクセスする。さらに、文書163でも、リンク元表記が「プレゼントはこちら」のリンク263があるために欲しい情報を得ようリンク263をたどる。ところが、リンク263の先は文書161に戻ってしまい、結局どこへ行くべきかわからなくなってしまう。このように、ループリンクがあると、訪問者は自分の欲しい情報を得られないまま、文書間をさ

迷うことになってしまう。

【００２３】

ハイパーテキストを検査する第1の従来技術としては、後述する非特許文献1で紹介されている、インターネット上のハイパーテキストを対象としたリンクチェッカーが挙げられる。これは、インターネット上に置かれたハイパーテキストを自動巡回して、エラーが発生したらそのログを記録するツールである。このリンクチェッカーには、検査対象のアドレスを指定してオンラインで診断するタイプと、ハードディスク上の特定のフォルダを指定してオフラインで診断するタイプが存在する。

【００２４】

また、第2の従来技術として、後述する特許文献1に示される発明が挙げられる。この方法によれば、管理すべきハイパーテキストのアドレスをデータベースに記憶しておき、そのアドレスに対してブラウザを定期的に自動接続させる。これにより記憶されたハイパーテキストのアドレスに文書が存在するか否かをチェックし、デッドリンクなどの物理的不整合を検知することが可能である。また、この発明によれば、データベース中の各文書を特定する手がかりとなるキーワードや画像をあらかじめシステムに登録する。これにより、デッドリンクを検知した場合には、検索エンジンを使って無くなったページを探し出し、訂正候補を提示することができる。

【００２５】

また、文書一般の検査の従来技術として、Microsoft社のWordのオートコレクト機能のような文書校正システムが挙げられる。これらの文書校正システムでは、送り仮名の間違いや助詞「の」の繰り返しなど、不適切な表現を検出し、訂正候補を出力することができる。

【００２６】

【特許文献1】

特開2001-273185号公報

【非特許文献1】

米国エルソプ(Elsop)社製のリンクチェッカー「LinkScan」、[平成14年10

月9日検索]、インターネット＜URL:http:/www.elsop.com/linkscan/＞

【００２７】

【発明が解決しようとする課題】

第1の問題点は、第1および第2の従来技術で検知できるのは物理的不整合のみであって、論理的不整合は検知できないことである。その理由は、第1および第2の従来技術ではハイパーテキストのアドレスに接続した際に、サーバーからのエラーが返ってくるか否かでしか、不整合の有無を判断していないからである。サーバーでエラーが発生しない論理的不整合の検知は、現状では人手によるブラウザ上での目視確認に頼るしか方法がない。

【００２８】

第2の問題点は、第1および第2の従来技術では物理的不整合の訂正候補しか提示できず、論理的不整合の訂正候補は提示できないことである。その理由は、第1の問題点と同じである。

【００２９】

第3の問題点は、人手によるブラウザ上の目視確認では、コストがかかり過ぎることである。その理由は、企業などの大規模なサイトは数千～数万件のハイパーテキストで構成されており、文書間のリンクは数万件～数十万件にもなるからである。これらのリンクをくまなくひとつづつ確認するのは時間の面でも費用の面でも現実的ではない。また、ブラウザ上の目視確認では幽霊リンクなどのチェックは漏れが生じやすい。

【００３０】

第4の問題点は、第3の従来技術では、リンク元表記の不統一のように、一つの文書へ異なる表現でリンク元表記が記述されているために閲覧者が混乱を生じる場合であっても、それを検出できないことである。その理由は、個々のリンク元表記が不適切な表現を含んでいなければ、正常と判断されるからである。

【００３１】

本発明の第1の目的は、物理的不整合だけでなく論理的不整合をも検知できるようにすることである。

【００３２】

本発明の第2の目的は、物理的不整合だけでなく論理的不整合の訂正候補をも管理者に提示できるようにすることである。

【００３３】

本発明の第3の目的は、不整合チェックのコストを大幅に削減することである。

【００３４】

【課題を解決するための手段】

本発明のハイパーテキスト検査装置は、ハイパーテキストデータベースを対象とし、リンク元表記とリンク先ページ内容との間に不整合が生じている箇所、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所、同一ページ内および周辺ページ内の複数のリンク元表記の間にスタイルの不統一が生じている箇所、リンク元表記のないリンク箇所、ループを形成するリンクの系列であって且つ該リンクの系列に対応するリンク元表記がすべて同一トピックに関わる箇所、のうちの少なくとも１種類の論理的なリンク不整合の箇所を検出することを基本とする。より具体的には以下のような構成を有する。

【００３５】

第１のハイパーテキスト検査装置は、ハイパーテキストを構成するページおよびリンクに関する情報を格納する情報記憶手段と、前記情報記憶手段を参照して論理的なリンク不整合の箇所を検出する条件判定手段とを備える。

【００３６】

第２のハイパーテキスト検査装置は、ハイパーテキストを構成するページおよびリンクに関する情報を収集する情報収集手段と、前記ページおよびリンクに関する情報を格納する情報記憶手段と、前記情報記憶手段を参照して論理的なリンク不整合の箇所を検出する条件判定手段とを備える。

【００３７】

第３のハイパーテキスト検査装置は、第１または第２のハイパーテキスト検査装置の構成に加えて、前記条件判定手段によって検出された箇所に関する訂正候補を計算する候補計算手段を備える。

【００３８】

第４のハイパーテキスト検査装置は、第３のハイパーテキスト検査装置の構成に加えて、前記条件判定手段によって検出された箇所の重要度を出力する重要度計算手段を備える。

【００３９】

第５のハイパーテキスト検査装置は、第３または第４のハイパーテキスト検査装置の構成に加えて、前記条件判定手段が検出したリンク不整合の箇所と前記候補計算手段が計算した訂正候補とに基づいて前記ハイパーテキストを更新する訂正反映手段を備える。

【００４０】

第６のハイパーテキスト検査装置は、第４のハイパーテキスト検査装置の構成に加えて、前記重要度計算手段で計算された重要度、前記条件判定手段によって検出された箇所の数、総リンク数に対する前記条件判定手段によって検出された箇所の数の割合、のうちの１ファクタもしくは複数のファクタの組み合わせによって前記ハイパーテキストに関するトータルスコアを計算して出力するトータルスコア計算手段を備える。

【００４１】

第７のハイパーテキスト検査装置は、第１または第２のハイパーテキスト検査装置の構成に加えて、前記条件判定手段によって検出された箇所の重要度を出力する重要度計算手段を備える。

【００４２】

第８のハイパーテキスト検査装置は、第７のハイパーテキスト検査装置の構成に加えて、前記重要度計算手段で計算された重要度、前記条件判定手段によって検出された箇所の数、総リンク数に対する前記条件判定手段によって検出された箇所の数の割合、のうちの１ファクタもしくは複数のファクタの組み合わせによって前記ハイパーテキストに関するトータルスコアを計算して出力するトータルスコア計算手段とを備える。

【００４３】

第１、２、７または８のハイパーテキスト検査装置において、前記条件判定手段

は、リンクに関する情報を特定の条件でグループ化し、グループから外れた特異なリンクに関する情報をリンク不整合の箇所として検出するものであって良い。

【００４４】

第１、２、７または８のハイパーテキスト検査装置において、前記条件判定手段は、リンク元表記とリンク先ページ内容との間に不整合が生じている箇所を検出するものであって良い。この場合、前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第２の不適正スコア、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第３の不適正スコア、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算される第４の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するものであって良い。

【００４５】

第１、２、７または８のハイパーテキスト検査装置において、前記条件判定手段は、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するものであって良い。この場合、前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク先ページ内容に含まれる移動や期限切れの告知表現を検出することで計算される第２の不適正スコア、または、（３）リンク元表記あるいはリンク先ページ内容に含まれる有効期間の記述と現在日時を比較することで計算される第３の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するものであって良い。

【００４６】

第１、２、７または８のハイパーテキスト検査装置において、前記条件判定手段は、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所を検出するものであって良い。

【００４７】

第１、２、７または８のハイパーテキスト検査装置において、前記条件判定手段は、同一ページ内および周辺ページ内の複数のリンク元表記の間にスタイルの不統一が生じている箇所を検出するものであって良い。

【００４８】

第３乃至第６のハイパーテキスト検査装置において、前記条件判定手段は、リンクに関する情報を特定の条件でグループ化し、グループから外れた特異なリンクに関する情報をリンク不整合の箇所として検出するものであって良く、また前記候補計算手段は、前記特異なリンクに関する情報を、不整合でない他のリンクと同一にするような訂正候補を求めるものであって良い。

【００４９】

第３乃至第６のハイパーテキスト検査装置において、前記条件判定手段は、リンク元表記とリンク先ページ内容との間に不整合が生じている箇所を検出するものであって良い。この場合、前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第２の不適正スコア、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第３の不適正スコア、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算される第４の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにし、前記候補計算手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算されるリンク元表記の訂正候補、または、（２）リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算されるリンク先の訂正候補、または、（３）リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算されるリンク先の訂正候補、または、（４）リンク元表記とリンク先ページとの間の内容比較によって計算されるリンク元表記の訂正候補、のうちのいずれか１種類もしくは複数種類を訂正候補とするものであって良い。

【００５０】

第３乃至第６のハイパーテキスト検査装置において、前記条件判定手段は、リンク先ページ内容が変更されてリンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するものであって良い。この場合、前記条件判定手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第１の不適正スコア、または、（２）リンク先ページ内容に含まれる移動や期限切れの告知表現を検出することで計算される第２の不適正スコア、または、（３）リンク元表記あるいはリンク先ページ内容に含まれる有効期間の記述と現在日時を比較することで計算される第３の不適正スコア、のうちのいずれか１種類もしくは複数種類を用いてリンクの不適正スコアを計算し、該不適正スコアの高い箇所を検出するようにし、前記候補計算手段は、（１）リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算されるリンク元表記の訂正候補、または、（２）リンク先ページ内容から移動先に関する情報を抽出することで計算されるリンク先の訂正候補、のうちの一方もしくは両方を訂正候補とするものであって良い。

【００５１】

第３乃至第６のハイパーテキスト検査装置において、前記条件判定手段は、同一のリンク先ページをもつ複数のリンク元表記の間に不統一が生じている箇所を検出するようにし、前記候補計算手段は、検出された箇所とリンク先ページが同一の複数リンクについてリンク元表記を比較することでリンク元表記の訂正候補を計算するものであって良い。

【００５２】

第３乃至第６のハイパーテキスト検査装置において、前記条件判定手段は、同一ページ内または周辺ページ内の複数のリンク元表記の間にスタイルの不統一が生じている箇所を検出するようにし、前記候補計算手段は、検出された箇所を含むページおよび周辺ページ内の複数のリンク元表記のスタイルを比較することでリンク元表記のスタイルの訂正候補を計算するものであって良い。

【００５３】

第２乃至第６のハイパーテキスト検査装置において、前記情報収集手段は、ハイ

パーテキストを構成するページおよびリンクに関する情報の収集を繰り返し実行し、前記情報格納手段は、複数の時刻における前記ページおよびリンクに関する情報を格納するものであって良い。この場合、前記条件判定手段は、前記情報格納手段を参照し、内容に変更が生じたページに対する被リンク数の時間変化やリンク元表記の種類の時間変化を計算することで、リンク元表記とリンク先ページ内容との間に不整合が生じた箇所を検出するものであって良い。

【００５４】

第１乃至第８のハイパーテキスト検査装置において、前記条件判定手段は、リンク元表記のないリンクを検出するものであって良い。

【００５５】

第１乃至第８のハイパーテキスト検査装置において、前記条件判定手段は、リンク元表記として文字列も画像も設定されていないリンク、あるいは、リンク元表記として目に付きにくい色やサイズを用いた文字列や画像が設定されたリンクを検出するものであって良い。

【００５６】

第１乃至第８のハイパーテキスト検査装置において、前記条件判定手段は、ループを形成するリンクの系列であって、かつ、該リンクの系列に対応するリンク元表記がすべて同一トピックに関わるものを検出するものであって良い。

【００５７】

第４または７のハイパーテキスト検査装置において、前記重要度計算手段は、（１）検出された箇所の誤り／不適切の種類、（２）検出された箇所の誤り／不適切の確度、（３）検出された箇所を含むページの被リンク数、（４）検出された箇所を含むページに対するユーザからのアクセス実績、（５）検出された箇所を含むページのハイパーテキストにおける階層レベル、のうちの１ファクタもしくは複数のファクタの組み合わせによって重要度を計算するものであって良く、また、検出された箇所の重要度を計算し、該重要度に基づいて、前記検出された箇所の出力件数や出力方法を制御するものであって良い。

【００５８】

第２乃至第８のハイパーテキスト検査装置において、前記情報収集手段は、リン

ク元表記が画像の場合に文字認識を行うことによって該リンク元表記に対応する文字列も抽出し、前記ページおよびリンクに関する情報の１つとして前記情報格納手段に登録するものであって良い。

【００５９】

第１乃至第８のハイパーテキスト検査装置は、Webサイトに存在するハイパーテキストを検査対象とするものであって良い。

【００６０】

他方、本発明の第１のハイパーテキスト検査方法は、ハイパーテキストデータベースを対象に、リンク元表記やリンク関係における誤り箇所あるいは不適切箇所を検出する条件を指定させるステップと、該条件に合致する箇所について（１）リンク元表記、（２）リンク元ページの識別情報、（３）リンク先ページの識別情報、という３項目を一覧表示するステップとを含んで構成される。ここで、前記一覧表示するステップにおいて、条件に合致した箇所を、（１）リンク元表記、（２）リンク元ページの識別情報、（３）リンク先ページの識別情報、という３項目のいずれかをソートキーとして一覧表示するものであって良い。また、前記３項目をスクリーンに一覧表示し、前記スクリーン上で（１）または（２）または（３）の項目を訂正させるステップと、スクリーン上で訂正された結果にしたがって前記ハイパーテキストデータベースを一括修正するステップとを含んでいて良い。更に、検査対象とするハイパーテキストデータベースを指定するステップを含んでいて良い。

【００６１】

また第２のハイパーテキスト検査方法は、Webサイトを構成するページおよびリンクに関する情報を収集する情報収集ステップと、前記情報収集ステップの結果を参照して論理的なリンク不整合の箇所を検出する条件判定ステップと、前記条件判定ステップによって検出された箇所の重要度およびWebサイトのトータルスコアを計算する重要度計算ステップとを、指定されたWebサイトを対象として定期的に自動実行し、前記指定されたWebサイトに関する前記トータルスコアの時間変化を通知する構成を有する。

【００６２】

また第３のハイパーテキスト検査方法は、Webサイトを構成するページおよびリンクに関する情報を収集する情報収集ステップと、前記情報収集ステップの結果を参照して論理的なリンク不整合の箇所を検出する条件判定ステップと、前記条件判定ステップによって検出された箇所の重要度およびWebサイトのトータルスコアを計算する重要度計算ステップとを、指定されたWebサイトを対象として定期的に自動実行し、前記指定されたWebサイトに関する前記トータルスコアあるいは前記検出された箇所の重要度があらかじめ定めた条件を満たした場合にアラートを通知する構成を有する。

【００６３】

また第４のハイパーテキスト検査方法は、Webサイトを構成するページおよびリンクに関する情報を収集する情報収集ステップと、前記情報収集ステップの結果を参照して論理的なリンク不整合の箇所を検出する条件判定ステップと、前記条件判定ステップによって検出された箇所の重要度およびWebサイトのトータルスコアを計算する重要度計算ステップとを、指定された複数のWebサイトを対象として自動実行し、前記指定された複数のWebサイトの間の前記トータルスコアの順位付け結果を通知する構成を有する。

【００６４】

【作用】

第１乃至第８のハイパーテキスト検査装置にあっては、特定の条件に該当するリンク情報をグループ化し、グループから外れた特異なリンクをリンク不整合として検出する等の処理によって、条件判定手段が論理的なリンク不整合の箇所を検出することにより、本発明の第1の目的を達成することができる。

【００６５】

第３乃至第６のハイパーテキスト検査装置にあっては、リンク不整合と不整合でない他のリンクのリンク情報の違いを基に、候補計算手段が、特異なリンクのリンク情報を、不整合でない他の大多数のリンクと同一にするような訂正候補を求める等の処理を行うことにより、本発明の第2の目的を達成することができる。

【００６６】

第１乃至第８のハイパーテキスト検査装置にあっては条件判定手段によって論理

的不整合が自動的に検出され、また、第３乃至第６のハイパーテキスト検査装置にあってはさらに候補計算手段によって訂正候補が自動的に求められ、また、第５のハイパーテキスト検査装置にあっては訂正反映手段によって論理的不整合箇所が自動的に訂正されることより、本発明の第3の目的を達成することができる。

【００６７】

【発明の第1の実施の形態】

次に、本発明の第1の実施の形態について、図面を参照して詳細に説明する。

【００６８】

図1を参照すると、本発明の第1の実施の形態は、プログラム制御により動作するデータ処理装置1と、情報を記憶する記憶装置2と、キーボード等の入力装置3と、ディスプレイ装置や印刷装置等の出力装置4とを備える。

【００６９】

データ処理装置1は、情報収集手段11、候補計算手段12、条件判定手段13および訂正反映手段14を備えている。

【００７０】

情報収集手段11は、記憶装置2に記憶されたハイパーテキストデータベース21から各文書を取り出し、リンク情報を取り出して情報記憶部22に格納する。ここでリンク情報は、リンク元アドレス、リンク先アドレス、リンク元表記、target属性、style属性などを含んでいる。なお、情報記憶部22には、リンク情報の他に、文書の本文、更新日付、取得日時、取得時の状態（エラーか成功かなど）を記録してもよい。

【００７１】

条件判定手段13は、情報記憶部22に格納されたリンクを各リンク情報の項目毎にグループ化し、グループから外れた特異なリンクをリンク不整合として情報記憶部22から抽出する。

【００７２】

候補計算手段12は、条件判定手段13が抽出したリンク不整合のリンクに対して、訂正候補を計算し出力する。ここで訂正候補では、不整合を起こしているリンク

のリンク情報のうち、どの項目をどのように訂正すべきかが指定される。

【００７３】

訂正反映手段14は、出力されたリンク不整合と訂正候補について、管理者が確認した結果をハイパーテキストデータベース21に反映させる。

【００７４】

記憶装置2は、ハイパーテキストデータベース21と情報記憶部22とを備えている。

【００７５】

ハイパーテキストデータベース21には、検査対象とするサイトに存在するハイパーテキストの集合が格納されている。なお、ハイパーテキストデータベース21は、必ずしもすべてがローカルな記憶装置2の中に存在している必要はなく、インターネット上のハイパーテキスト群のようにネットワークを介して分散していてもよい。

【００７６】

情報記憶部22には、ハイパーテキストデータベース21中の各文書に含まれるリンク情報が格納されている。例えば、図2の文書101に含まれるリンク情報は図9のようになる。図9を見ると、リンク201は文書101から文書102へリンク元表記「GX0011」でリンクされており、target属性は_blank、style属性はst01と指定されていることが分かる。なお、ここではリンク元表記がテキストの場合について説明したが、リンク元表記に画像が指定されている場合は、指定された画像ファイルのアドレスをリンク元表記に記録する。また、画像ファイルを文字認識モジュールにかけて、画像内部に記述されているテキストを抽出し文字列と同様に登録してもよい。

【００７７】

次に、図1、図9～13を参照して本実施の形態の動作について詳細に説明する。

【００７８】

まず、入力手段3から入力された収集条件の設定に基づき、情報収集手段11がハイパーテキストデータベース21に格納されている文書を読み出す（図10のステップS1）。ここで、ハイパーテキストデータベース21がWWWの場合、HTTP（Hyper T

ext Transfer Protocol）を介して文書にアクセスすることができる。このような機能は、従来、IE（Internet Explorer)などのWebブラウザ、あるいはWebクローラー（スパイダー/ロボット）において実現されている。ハイパーテキストデータベース21がWWWの場合の収集設定画面を図11に示す。図11では、分析対象とするサイトのドメイン名、収集する文書の目標ページ数、収集対象とする文書の拡張子、サーバーにアクセスする時間間隔、収集に失敗した場合のリトライ回数、収集時のタイムアウト時間、リンクをたどって再帰的に収集する場合の再帰の階層の深さなどを指定できる。図11の実行ボタンを押すと、ハイパーテキストの収集が開始される。

【００７９】

次に、情報収集手段11は収集した文書のHTML記述を解析し、図9に示すようなリンク情報を抽出して情報記憶部22に格納する（図10のステップS2）。

【００８０】

次に、入力手段3から入力された抽出条件に基づき、条件判定手段13が該当するリンクを情報記憶部22からリンク不整合として抽出する（図10のステップS3）。抽出条件の設定画面を図12に示す。図12では、分析対象となるサイトについて、デッドリンク（物理的不整合）、間違いリンク、期限切れ情報へのリンクなど、各種のリンク不整合のうち、どれを抽出するかを指定することができる。また、あらかじめ特定のアドレスへのリンクが不整合であると分かっている場合、そのアドレスをリンク先に持つリンクを抽出することもできる（図12中の「特定URL」）。さらに、抽出されるリンク不整合が多い場合に、何件づつ画面に表示するかも指定可能である。図12の実行ボタンを押すと、リンク不整合の抽出が開始される。各リンク不整合のうち、デッドリンクの抽出は前述した従来技術によって可能であり、本発明とは直接関係しないので説明は省略する。また、特定URLをリンク先に持つリンクの抽出方法は当業者に自明であるためその説明は省略する。残りの各種論理的不整合リンクの抽出方法の詳細は後述する。

【００８１】

次に、条件判定手段13がリンク不整合として抽出したリンクについて、候補計算手段12が不整合を解消するための訂正候補を求め、結果一覧画面を出力する（図

10のステップS4、S5）。

【００８２】

出力されるリンク不整合の結果一覧画面の一例を図13に示す。図13ではリンク先とリンク元表記が同じリンクをグループ化しており、それぞれ不整合の種類、訂正候補を付与して表示している。また、各リンク元アドレスおよびリンク先アドレスをクリックすると該当文書にアクセスできるようになっている。また、訂正候補の欄にはシステムが出力する訂正候補が記入されている。訂正候補は訂正対象とすべきリンク情報の項目と、どのように訂正すべきかを「：」で区切って記述する。例えば、図13で「リンク：削除」とあるのは、リンク自体を削除することを意味する。また、「リンク元表記：新着情報」とあるのは、リンク元表記を「新着情報」に変更することを意味する。この訂正候補は、管理者が確認後、書き換えることも可能である。

【００８３】

次に、管理者は出力されたリンク不整合と訂正候補を確認する（図10のステップS6）。このとき、図13ではリンク先とリンク元表記が同じリンクをグループ化しているため、管理者はリンクをすべて確認しなくても、各不整合の代表例のみを確認すればよい。例えばリンク271～274のリンクはどれもリンク先が文書175で、リンク元表記が「○×キャンペーン実施中」で、期限切れになっており、それを訂正するためには、リンクを削除する必要があることがわかる。そのため、管理者は、リンク271～274のリンクをすべて確認しなくとも、文書171にアクセスしてリンク271の不整合と訂正候補が正しいと確認できれば、残りのリンク272～274を確認する必要はなく、確認に要するコストが削減される。

【００８４】

訂正候補が複数ある場合は、「リンク先：文書177 OR リンク元表記：製品B」のようにORで区切って管理者に提示される。この場合、管理者は、確認の結果必要な訂正候補のみを残せばよい。また、確認の結果、訂正候補が間違っていると判断した場合、それを修正することもできる。例えば、リンク278、279の訂正候補はリンク元表記を「新着情報」に訂正するようになっているが、リンク先アドレスを文書180に変更した方が適当だと考えた場合には、該当する訂正候補を「リ

ンク先：文書180」に変更すればよい。また、管理者は、確認の結果、訂正したくないと判断した場合には、訂正候補を空欄にすれば後のステップで訂正は行われない。

【００８５】

次に、管理者が図13の訂正反映ボタンを押すと、訂正反映手段15は、管理者に確認された訂正候補に基づいてハイパーテキストデータベース21の各文書を修正する（図10のステップS7）。この段階で、訂正候補が複数ORでつながれている場合は、最初の訂正候補だけが反映される。

【００８６】

また、図13では、リンク元、リンク先、リンク元表記の項目に「ソート」というリンクがある。これはそれぞれの項目をキーに抽出結果をソートするためのものである。例えば、リンク元の項目にある「ソート」をクリックすると、リンク元文書をキーに抽出結果をソートして出力する。これにより、各文書にそれぞれどのようなリンク不整合が発生しているかを把握することができるため、不整合を人手で修正する場合に利用可能である。また、リンク先の項目にある「ソート」をクリックすると、リンク先文書をキーに抽出結果をソートして出力する。これにより、ある特定の文書にはられたリンクについて、不整合の発生状況を把握することができるため、アクセスが集中する文書など、重要な文書に対する不整合を重点的に調べることができる。さらに、リンク元表記の項目にある「ソート」をクリックすると、リンク元表記をキーに抽出結果をソートして出力する。これにより、どのような種類のリンク元表記において不整合が発生しやすいかを把握することができるため、リンク元表記として使っている表現の妥当性などを調べることができる。

【００８７】

なお、本実施の形態では、図13の結果一覧画面における「訂正候補」の欄に表示されたリンク元表記、リンク先等を管理者に訂正させたが、同画面における「リンク元」、「リンク先」、「リンク元表記」の欄を直接上書きすることで、リンク元、リンク先、リンク元表記を訂正させるようにしても良い。また、本実施の形態では、ハイパーテキストの収集設定と、リンク不整合の抽出条件設定を別々

の画面で行ったが、分析を開始する時点で同じ画面で一度に条件を設定しておき、ステップS1～S5までをすべて自動化して実行する方法もあり、本実施の形態に述べた方法に限定されない。

【００８８】

また、本実施の形態では、ステップS6で管理者が出力されたリンク不整合と訂正候補の確認を行ったが、ステップS6を省略してステップS1～S7までをすべて自動化して実行する方法もあり、本実施の形態に述べた方法に限定されない。

【００８９】

また、本実施の形態では、管理者が検査のタイミングを決めて実行する場合について説明したが、あらかじめ収集条件と抽出条件を設定しておき、定期的に自動でステップS1～S5までを実行し、得られた結果をメールなどで通知する方法などもあり、本実施の形態に述べた方法に限定されない。

【００９０】

[間違いリンク検知の実施形態]

次に、図3および図14、図15を参照して、間違いリンクを検知する場合の条件判定手段13と候補計算手段12の動作について詳細に説明する。情報記憶部22には、図3の文書群のリンク情報が格納されているものとする。

【００９１】

まず、条件判定手段13は情報記憶部22から、リンク元表記が同じリンクをグループ化し、同一グループ内でリンク先が同じリンクをサブグループ化し、リンク先が異なるサブグループに属するリンクを抽出する。また、サブグループに含まれるリンクの数に応じて、各リンクに不適正スコアを付与する（ステップT11）。図15(a)に、ステップT11で抽出されるリンクと、付与される不適正スコアの例を示す。図15(a)を見ると、リンク元表記「GX0011」でリンク211、212、213、214がグループ化されており、リンク元表記「GX0012」でリンク215、216がグループ化されていることがわかる。さらに、リンク元表記「GX0011」のグループのうち、リンク211、212、213の3件はリンク先が文書116のサブグループになり、リンク214はリンク先が文書117のサブグループとなる。また、リンク元表記「GX0012」のグループのうち、リンク215はリンク先が文書116のサブグループ、リンク21

6はリンク先が文書117のサブグループとなる。

【００９２】

不適正スコアの付与は、まず一つのグループの不適正スコアを1とし、それをサブグループ内のリンク数に反比例して配分したものを各サブグループの不適正スコアとする。さらに、各サブグループの不適正スコアをサブグループ内のリンクの数で等分したものを各リンクの不適正スコアとする。例えば、図15(a)では、リンク元表記「GX0011」のグループの不適正スコアを1とし、サブグループ内のリンク数に反比例して配分すると、リンク先アドレスが文書116のサブグループの不適正スコアは1/4、リンク先アドレスが文書117のサブグループの不適正スコアは3/4になる。さらに、リンク211、212、213でサブグループの不適正スコア1/4を3等分するため、各リンクの不適正スコアは1/12となる。また、リンク214の不適正スコアは3/4である。同様に、リンク215、216の不適正スコアはどちらも1/2となる。

【００９３】

次に、条件判定手段13は情報記憶部22から、リンク先が同じリンクをグループ化し、同一グループ内でリンク元表記が同じリンクをサブグループ化し、リンク元表記が異なるサブグループに属するリンクを抽出する。また、サブグループに含まれるリンクの数に応じて、各リンクに不適正スコアを付与する（ステップT12）。図15(b)に、ステップT12で抽出されるリンクと、付与される不適正スコアの例を示す。図15(b)を見ると、リンク先が文書116のリンク211、212、213、215がグループ化されており、リンク先が文書117のリンク214、216がグループ化されていることがわかる。さらに、リンク先が文書116のグループのうち、リンク211、212、213の3件はリンク元表記が「GX0011」のサブグループになり、リンク215はリンク元表記が「GX0012」のサブグループとなる。また、リンク先が文書117のグループのうち、リンク214はリンク元表記が「GX0011」のサブグループ、リンク216はリンク元表記が「GX0012」のサブグループとなる。不適正スコアの付与は、ステップT11と同じである。したがって、ステップT12でのリンク211、212、213の不適正スコアは1/12、リンク215の不適正スコアは3/4、リンク214、216の不適正スコアは1/2となる。

【００９４】

次に、条件判定手段13は情報記憶部22から、リンク元が同じでかつリンク元表記も同じリンクをグループ化し、同一グループ内でリンク先が同じリンクをサブグループ化し、リンク先が異なるサブグループに属するリンクを抽出する。また、サブグループに含まれるリンクの数に応じて、各リンクに不適正スコアを付与する（ステップT13）。図15(c)に、ステップT13で抽出されるリンクと、付与される不適正スコアの例を示す。図15(c)を見ると、リンク元が文書115でかつリンク元表記が「GX0012」であるリンク215、216がグループ化されていることがわかる。さらに、リンク215はリンク先が文書116のサブグループになり、リンク216はリンク先が文書117のサブグループとなる。不適正スコアの付与は、ステップT11と同じである。したがって、ステップT13でのリンク215、216の不適正スコアはそれぞれ1/2となる。

【００９５】

次に、条件判定手段13は情報記憶部22から、リンク元表記に含まれる単語がリンク先文書のタイトル、見出し、強調文字列に含まれないリンクを抽出し、不適正スコア1を付与する（ステップT14）。図15(d)に、ステップT14で抽出されるリンクと、付与される不適正スコアの例を示す。図15(d)で抽出されているリンク214、215は、図3において、どちらもリンク先の文書にリンク元表記に含まれる単語が出現していない。

【００９６】

次に、条件判定手段13は、各リンクの不適正スコアを合計する（ステップT15）。したがって、リンク211、212、213の不適正スコアは1/12+1/12=1/6となる。また、リンク214の不適正スコアは3/4+1/2+1=9/4となる。また、リンク215の不適正スコアは1/2+3/4+1/2+1=11/4となる。また、リンク216の不適正スコアは1/2+1/2+1/2=3/2となる。

【００９７】

次に、条件判定手段13は、各サブグループ間で不適正スコアの合計を比較し、不適正スコアが高いリンクをリンク不整合として抽出する。また、候補計算手段12は、各条件で抽出されたリンクについて、同一グループ内で、スコアの高いリン

クのリンク情報を、スコアの低いリンクのリンク情報に一致させるような訂正候補を求める（ステップT16）。図15(a)の、リンク元表記が「GX0011」であるグループでは、リンク211、212、213からなるサブグループの不適正スコアの合計は1/6+1/6+1/6=1/2、リンク214からなるサブグループの不適正スコアの合計は9/4であるから、不適正スコアが高いリンク214をリンク不整合と決定する。また、リンク214のリンク情報をリンク211、212、213のサブグループに一致させるためには、「リンク先：文書116」の訂正候補が適当であることがわかる。さらに、図15(a)の、リンク元表記が「GX0012」であるグループでは、リンク215の不適正スコアの合計は11/4、リンク216の不適正スコアの合計は3/2であるため、リンク215をリンク不整合と決定する。また、リンク215のリンク情報をリンク216のサブグループに一致させるためには、「リンク先：文書117」の訂正候補が適当であることがわかる。同様に、図15(b)では、リンク215がリンク不整合と決定されて「リンク元表記：GX0011」が訂正候補となり、リンク214がリンク不整合と決定されて「リンク元表記：GX0012」が訂正候補として求まる。さらに同様に、図15(c)では、リンク215がリンク不整合と決定されて「リンク先：文書117」が訂正候補となる。以上の結果を統合すると、リンク不整合はリンク214、215であり、訂正候補はそれぞれ「リンク先：文書116 OR リンク元表記：GX0012」、「リンク先：文書117 OR リンク元表記：GX0011」となる。

【００９８】

なお、本実施の形態では、不適正スコアの合計が高いリンクをリンク不整合としたが、不適正スコアの閾値を設けて、不適正スコアが高くても閾値以下であればリンク不整合としない方法もあり、本実施の形態に述べた方法に限定されない。

【００９９】

また、本実施の形態では、不適正スコアの一例としてサブグループ内でのリンク数を元に計算したが、単純に抽出された回数を不適正スコアとしてもよく、本実施の形態に述べた方法に限定されない。また、サブグループ内でのリンク数をそのリンクの特徴ベクトルとし、あらかじめ教師データとして与えられた不整合リンクの特徴ベクトルとの距離の平均値を不適切スコアとする方法などもあり、本実施の形態に述べた方法に限定されない。

【０１００】

また、本実施の形態では、間違いリンクの抽出条件として、(1)リンク先ページが同一の複数リンクについてリンク元表記を比較することで計算される第1の不適正スコア、(2)リンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第2の不適正スコア、(3)リンク元ページおよびリンク元表記が同一の複数リンクについてリンク先ページを比較することで計算される第3の不適正スコア、(4)リンク元表記とリンク先ページとの間の内容比較によって計算される第4の不適正スコアを合計して求めたが、これらのうち、1種類もしくは複数種類を用いたり、各条件に応じて重み付けを行って不適正スコアを計算してもよく、本実施の形態に述べた方法に限定されない。

【０１０１】

[期限切れリンク検知の実施形態]

次に、図4および図16を参照して、期限切れリンクを検知する場合の条件判定手段13と候補計算手段12の動作について詳細に説明する。

【０１０２】

まず、条件判定手段13はリンク元表記に日付表現が含まれるリンクや、日付表現が含まれる文書を指しているリンクを抽出して日付表現から有効期限を計算し、現在日時が有効期限内であるか否かを判定する（図16のステップT21）。

【０１０３】

次に、条件判定手段13は抽出したリンクのリンク先文書に含まれる期限切れ表現を抽出する（図16のステップT22）。ここで、期限切れ表現とは、「閉鎖しました」「移動しました」「終了しました」「秒後に自動的にジャンプします」「○月×日をもちまして」「ご愛顧ありがとうございました」「参加ありがとうございました」など、サービスが終了、閉鎖または移動した場合の告知文によく使われる表現のことである。また、上記表現の他にも、HTMLによって数秒後に自動的に文書が切り替わる設定になっていればこれも期限切れ表現として抽出する。

【０１０４】

次に、条件判定手段13は、ステップT21の有効期限内か否かの判定結果と、ステップT22で抽出された期限切れ表現の数を統合して、リンクの不適切スコアを計

算する。この不適切スコアがあらかじめ定めた閾値以下であれば、リンク不整合として出力する（図16ステップT23）。リンクの不適切スコアの計算方法の例としては、有効期限からの日数と、抽出された期限切れ表現の出現回数とを掛けて求める方法などがある。なお、その他にも、有効期限内か否かの結果と、抽出された期限切れ表現の出現回数をそのリンクの特徴ベクトルとし、あらかじめ教師データとして与えられた不整合リンクの特徴ベクトルとの距離の平均値を不適切スコアとする方法などもあり、本実施の形態に述べた方法に限定されない。

【０１０５】

次に、候補計算手段12は、リンク不整合として出力されたリンクについて、リンク先文書中から移転先アドレスを抽出して訂正候補とする。ここで、移転先のアドレスとは、HTMLによって自動的に文書が切り替わる設定になっている場合のとび先のアドレスである。また、自動的に文書が切り替わらなくても、「ここをクリック」「下記URLに移動しました」などの表現を抽出し、その表現の内部あるいは近傍に記述されているリンクのリンク先アドレスを移転先アドレスとして訂正候補にしてもよい。一方、移動先アドレスが抽出できなかった場合は、訂正候補を「リンク：削除」として出力する。

【０１０６】

図4(a)の場合について、条件判定手段13と候補計算手段12の動作の具体例を述べる。なお、リンクの不適切スコアの計算方法としては、前述したように、有効期限からの日数と、抽出された期限切れ表現の出現回数とを掛けて求める方法を用いるものとする。

【０１０７】

まず、ステップT21では、文書125内に「2002年7月20日～2002年8月31日」という日付表現があるため、条件判定手段13はリンク221、222、223、224を抽出する。この時、現在日時は2002年8月15日であるため、リンク221、222、223、224は有効期限内であると判定される。

【０１０８】

次に、ステップT22では、文書125には期限切れ表現は出現しないため、何も抽出されない。

【０１０９】

次に、ステップT23では、ステップT21で日付表現が有効期限内であった結果と、ステップT22で何も期限切れ表現が抽出されなかったことから、リンク221、222、223、224の不適切スコアは、有効期限からの日数および抽出された期限切れ表現の出現回数が共に0であるため0×0＝0であり、どのリンクも適切であると判定される。

【０１１０】

一方、図4(b)の場合について、条件判定手段13と候補計算手段12の動作の具体例を述べる。

【０１１１】

まず、ステップT21では、文書125内に「2002年7月20日～2002年8月31日」という日付表現があるため、条件判定手段13はリンク224を抽出する。この時、現在日時は2002年9月15日であるため、リンク224は文書T125内での有効期限を超えていると判定される。

【０１１２】

次に、ステップT22では、条件判定手段13は文書125から「終了しました」という期限切れ表現を抽出する。

【０１１３】

次に、ステップT23では、ステップT21で日付表現が有効期限外であった結果と、ステップT22で「終了しました」という期限切れ表現が抽出されたことから、リンク224の不適切スコアは、有効期限からの日数が15、抽出された期限切れ表現の出現回数が1であるため、15×1＝15である。ここで、例えば閾値が10であれば、リンク224はリンク不整合と判定される。

【０１１４】

次に、ステップT24では、候補計算手段12が移転先アドレスを抽出しようとするが、文書125には該当するアドレスが記述されていないため、「リンク：削除」をリンク224の訂正候補として出力する。

【０１１５】

なお、本実施の形態では、日付表現と期限切れ表現に注目した場合の期限切れリ

ンクの検知について説明したが、間違いリンクの検知と同様にリンク先ページが同一のリンクをグループ化し、同一グループ内でリンク元表記が異なるサブグループを検知する方法や、リンク元表記が同一のリンクをグループ化し、同一グループ内でリンク先が異なるサブグループを検知する方法などもあり、本実施の形態に述べた方法に限定されない。

【０１１６】

[リンク元表記の不統一検知の実施形態]

次に、図5および図17、図18を参照して、リンク元表記の不統一を検知する場合の条件判定手段13と候補計算手段12の動作について詳細に説明する。

【０１１７】

まず、条件判定手段13は情報記憶部22から、リンク先が同じリンクをグループ化し、同一グループ内でリンク元表記が同じリンクをサブグループ化し、リンク元表記が異なるサブグループに属するリンクを抽出する。また、サブグループに含まれるリンクの数に応じて、各リンクに不適正スコアを付与する（図17のステップT31）。文書群が図5の場合、ステップT31で抽出されるリンクと、付与される不適正スコアは図18のようになる。図18を見ると、リンク先が文書135のリンク231、232、233、234がグループ化されていることがわかる。さらに、リンク231、232、233の3件はリンク元表記が「GX Series」のサブグループになり、リンク234はリンク元表記が「gX Series」のサブグループとなる。

【０１１８】

不適正スコアの付与は、まず一つのグループの不適正スコアを1とし、それをサブグループ内のリンク数に反比例して配分したものを各サブグループの不適正スコアとする。さらに、各サブグループの不適正スコアをサブグループ内のリンクの数で等分したものを各リンクの不適正スコアとする。したがって、ステップT31でのリンク231、232、233の不適正スコアは1/12、リンク234の不適正スコアは3/4となる。ここで、条件判定手段13は、各サブグループ間で不適正スコアの合計を比較し、不適正スコアが高いリンクをリンク不整合として抽出する。図18では、リンク231、232、233の不適正スコアの合計1/4よりもリンク134の不適正スコア3/4が高いため、リンク134をリンク不整合として抽出する。

【０１１９】

次に、候補計算手段12は、抽出されたリンクのリンク元表記が用語辞書に登録されているか否かを調べる（図17のステップT32）。ここで用語辞書とは、ある単語について表記揺らぎをキーとして、統一すべき表現を値として持つテーブルである。例えば、「フリーソフトウェア」は無料で利用できるソフトウェアの意味で、「フリーウェア」「フリーソフト」などの表記揺れが存在するが、文書管理者のポリシーとしてすべて「フリーソフトウェア」に統一したい場合は、「フリーウェア」「フリーソフト」をキーに、「フリーソフトウェア」を値として用語辞書に登録しておけばよい。もし、抽出されたリンクのリンク元表記が用語辞書に登録されていた場合、候補計算手段12はキーに対応する統一すべき表現を訂正候補として出力する（図17のステップT33）。なお、表記ゆれを十分に吸収するためにキーの検索時に、あいまい検索を使用してもよい。また、表記揺らぎの単語を用いずに、統一すべき表現自身をあいまい検索し、文字列の類似度が閾値以上であれば、検索された統一すべき表現を訂正候補としてもよい。

【０１２０】

図18の場合、「GX Series」「gX Series」のいずれも用語辞書に登録されていなかったとする。

【０１２１】

一方、抽出されたリンクのリンク元表記が用語辞書に登録されていなかった場合、候補計算手段12は同一グループ内で、不適性スコアの大きいリンクのリンク元表記を、スコアの小さいリンクのリンク元表記に一致させるような訂正候補を求める（図17のステップT34）。図18の場合、「リンク元表記：GX Series」を訂正候補として出力する。

【０１２２】

なお、本実施の形態では、不適正スコアの一例としてサブグループ内でのリンク数を元に計算したが、サブグループ内でのリンク数をそのリンクの特徴ベクトルとし、あらかじめ教師データとして与えられた不整合リンクの特徴ベクトルとの距離の平均値を不適切スコアとする方法などもあり、本実施の形態に述べた方法に限定されない。

【０１２３】

[リンク元表記のスタイルの不統一検知の実施形態]

次に、図6および図19、図20を参照して、リンク元表記のスタイルの不統一を検知する場合の条件判定手段13と候補計算手段12の動作について詳細に説明する。

【０１２４】

まず、条件判定手段13は情報記憶部22から、リンク元文書が同一のリンクをグループ化し、同一グループ内でtarget属性が同じリンクをサブグループ化し、target属性が異なるサブグループに属するリンクを抽出する。また、サブグループに含まれるリンクの数に応じて、各リンクに不適正スコアを付与する（図19のステップT41）。文書群が図6の場合、ステップT41で抽出されるリンクと、付与される不適正スコアは図20のようになる。図20を見ると、リンク元が文書141のリンク241、242、243、244がグループ化されていることがわかる。さらに、リンク241、242、243の3件はtarget属性が「_blank」のサブグループになり、リンク244はtarget属性が無指定のサブグループとなる。

【０１２５】

不適正スコアの付与は、まず一つのグループの不適正スコアを1とし、それをサブグループ内のリンク数に反比例して配分したものを各サブグループの不適正スコアとする。さらに、各サブグループの不適正スコアをサブグループ内のリンクの数で等分したものを各リンクの不適正スコアとする。したがって、ステップT41でのリンク241、242、243の不適正スコアは1/12、リンク244の不適正スコアは3/4となる。ここで、条件判定手段13は、各サブグループ間で不適正スコアの合計を比較し、不適正スコアが高いリンクをリンク不整合として抽出する。図20では、リンク241、242、243の不適正スコアの合計1/4よりもリンク144の不適正スコア3/4が高いため、リンク144をリンク不整合として抽出する。

【０１２６】

次に、候補計算手段12は同一グループ内で、不適正スコアの大きいリンクのtarget属性を、スコアの小さいリンクのtarget属性に一致させるような訂正候補を求める（図19のステップT42）。図20の場合、「target属性：_blank」を訂正候補として出力する。

【０１２７】

なお、本実施の形態では、ステップT41でグループ化する対象をリンク元文書が同じリンクとしたが、リンク元文書が同じリンク群のうち、テーブルやリンクのリストなど、特定領域に存在するリンクに限ってグループ化する方法もあり、本実施の形態に述べた方法に限定されない。また、特定文書と同じディレクトリに格納されている文書など、複数の文書間でのリンクを、スタイルを基準にグループ化し、特定文書の周辺ページのリンクスタイルの不統一を検出する方法もあり、本実施の形態に述べた方法に限定されない。

【０１２８】

また、本実施の形態では、target属性の不統一の検知と訂正候補の求め方について述べたが、同様の方法でstyle属性の不統一の検知と訂正候補を求めることができる。

【０１２９】

また、本実施の形態では、不適正スコアの一例としてサブグループ内でのリンク数を元に計算したが、サブグループ内でのリンク数をそのリンクの特徴ベクトルとし、あらかじめ教師データとして与えられた不整合リンクの特徴ベクトルとの距離の平均値を不適切スコアとする方法などもあり、本実施の形態に述べた方法に限定されない。

【０１３０】

[幽霊リンク検知の実施形態]

次に、図7および図21を参照して、幽霊リンクを検知する場合の条件判定手段13と候補計算手段12の動作について詳細に説明する。

【０１３１】

まず、条件判定手段13は情報記憶部22から、不可視なリンク元表記が指定されているリンクを抽出する（図21のステップT51）。ここで、不可視なリンク元表記とは、空文字列、透明な画像、非常に小さな画像や文字、背景と同じ色の画像や文字などのことである。図7(a)では、リンク元表記に空文字列が指定されているリンクが抽出される。

【０１３２】

次に、候補計算手段12は、リンクを削除するよう訂正候補を「リンク：削除」として出力する（図21のステップT52）。

【０１３３】

[ループリンク検知の実施形態]

次に、図8および図22を参照して、ループリンクを検知する場合の条件判定手段13の動作について詳細に説明する。なお、候補計算手段12はループリンク検知時は動作しない。

【０１３４】

まず、条件判定手段13は、情報記憶部22に格納されているリンクの、リンク元表記を単語に分割する（図22のステップT61）。リンク元表記を単語に分割する方法としては、形態素解析を使う、字種の変わり目で切る、n文字毎に切るなどの方法がある。

【０１３５】

次に、条件判定手段13は、ループを形成するリンクの系列であって、かつ、該リンクの系列に対応するリンク元表記中の単語がすべて同一のリンク群を抽出する（図22のステップT62）。図8では、単語「プレゼント」を含むリンク261、262、263はループを構成しているため、ループリンクとして出力される。

【０１３６】

なお、本実施の形態では、リンク元表記中の単語がすべて同一のループリンクを抽出する場合について説明したが、トピック毎に特徴的な単語の辞書を持っておき、リンク元表記中の単語がすべて同一トピックに属するループリンクを抽出する方法もあり、本実施の形態に述べた方法に限定されない。

【０１３７】

[時間変化に注目したリンク不整合検知方法]

本実施の形態では、ある時点で収集した各リンクのリンク情報を基に各種のリンク不整合を検知する方法について述べたが、リンク情報の収集を定期的に繰り返し実行し、リンク情報の時系列変化に着目して各種リンク不整合を検知する方法もある。図4および図23、24を参照して、リンク情報の時系列変化に着目して各種リンク不整合を検知する場合の、条件判定手段13と候補計算手段12の動作につ

いて詳細に説明する。

【０１３８】

情報格納手段22には、時間Tと時間T'におけるリンク情報を格納しているものとする。

【０１３９】

まず、条件判定手段13は、時間Tと時間T'においてリンク情報の一項目が同一のリンクをグループ化する（図23のステップT71）。図4の場合、2002年8月15日時点でのリンク情報と、2002年9月15日時点でのリンク情報について、リンク先が文書125のリンクをグループ化すると、図24のようになる。

【０１４０】

次に、同一グループ内で多数のリンクのリンク情報が変化したリンクをリンク不整合として抽出する（図23のT72）。図23の場合、2002年8月15日の時点では、リンク先が文書125のリンクが4件あるのに対し、2002年9月15日時点では、リンク先が文書125のリンクは1件しかない。そこで、リンク224をリンク不整合として抽出する。

【０１４１】

次に、候補計算手段12は、時間Tと時間T'で起こった変化に対応する訂正候補を出力する（図23のステップT72）。図23の場合、2002年8月15日と2002年9月15日では、リンクの削除が起こっているので、「リンク：削除」を訂正候補として出力する。

【０１４２】

なお、本実施の形態では、時間Tと時間T'におけるリンク先文書が同一のリンクをグループ化した時に、リンクの削除が起こっている場合について述べたが、リンク元表記が変化している場合は、時間Tでのリンク先文書の内容が時間T'で変化したものとして、候補計算手段12はリンク元表記を変更するよう、訂正候補を出力する。

【０１４３】

また、本実施の形態では、時間Tと時間T'においてリンク先文書が同一のリンクをグループ化する方法について述べたが、他にもリンク元表記が同一のリンクを

グループ化して、style属性やtarget属性の変化を検知する方法などもあり、本実施の形態に述べた方法に限定されない。

【０１４４】

次に、本実施の形態の効果について説明する。

【０１４５】

本実施の形態では、各種の論理的不整合を検知することができる。すなわち、本実施の形態では、ハイパーテキストデータベースからリンク情報を抽出し、リンク情報の各項目毎にリンクをグループすることにより、グループから外れた特異なリンクをリンク不整合として検知するため、(1)リンクの張り間違い、(2)期限切れ情報へのリンク、(3)リンク元表記の不統一、(4)リンク元表記のスタイルの不統一といった論理的不整合を検知することができる。また、リンク情報の収集を定期的に繰り返し実行し、リンク情報の時系列変化に着目して各種リンク不整合を検知する方法によっても、(2)期限切れ情報へのリンクといった論理的不整合を検知することができる。更に、リンク元表記のないリンクの検出により、論理的不整合の一形態である(5)幽霊リンクも検知することができ、ループを形成するリンクの系列であって、かつ、該リンクの系列に対応するリンク元表記がすべて同一トピックに関わるものを検出することにより、論理的不整合の一形態である(6)ループリンクも検知することができる。

【０１４６】

また、本実施の形態では、論理的不整合の訂正候補を管理者に提示することができる。すなわち、本実施の形態では、グループから外れた特異なリンクのリンク情報を、グループと同一のリンク情報になるように訂正候補を自動計算する等の処理によって訂正候補を求めるため、管理者は不整合をどのように修正すべきかを検討する必要がなく、自動で修正を反映することも可能である。

【０１４７】

また、本実施の形態では、リンク不整合をグループ化してまとめて表示する。そのため、管理者は一部のリンクを確認すれば残りのリンクも同様に不整合か否かを判定でき、管理者によるチェックの効率が大幅に向上する。

【０１４８】

また、本実施の形態では、リンク不整合を（１）リンク元表記、（２）リンク元ページの識別情報、（３）リンク先ページの識別情報、という３項目のいずれかをソートキーとして一覧表示する。そのため、管理者は、ページ単位での修正項目を把握したり、重要なページに対する不整合を重点的に調べたり、リンク元表記として使っている表現の妥当性などを調べることができる。

【０１４９】

なお、本実施の形態のデータ処理装置１は情報収集手段11を備えているが、情報収集手段11によるハイパーテキストデータベース21からのページおよびリンクに関する情報の収集と記憶を別のデータ処理装置で実施する形態も考えられ、そのような形態ではデータ処理装置１の情報収集手段11は省略可能である。また、図13の結果一覧画面を見て管理者自身の手作業によってハイパーテキストデータベース21の不整合箇所を訂正する構成にあっては、訂正反映手段14を省略することができる。さらに、管理者の負担が増大するものの、図13の結果一覧画面の不整合の種類、訂正候補が無くても残りの情報から管理者自身で訂正候補を求めることもできるため、候補計算手段12を省略した構成も採用可能である。

【０１５０】

【発明の第2の実施の形態】

次に、本発明の第2の実施の形態について、図面を参照して詳細に説明する。

【０１５１】

図25を参照すると、本発明の第2の実施の形態は、データ処理装置5が、図1に示された第1の実施の形態におけるデータ処理装置2の構成に加えてさらに、重要度計算手段15を備えている点で第1の実施の形態と異なる。

【０１５２】

重要度計算手段15は、条件判定手段13が抽出したリンク不整合に対して、リンク不整合が検知された文書へのアクセス頻度や、不整合の深刻さに応じて重要度を計算しランキングして出力する。

【０１５３】

次に、本実施の形態の動作を図面を参照して詳細に説明する。

【０１５４】

図26のステップS1～S3で示される本実施の形態における情報収集手段11、条件判定手段13の動作は、第1の実施の形態の各手段11、13の動作と同一のため、説明は省略する。また、候補計算手段12は、条件判定手段13がリンク不整合として抽出したリンクについて不整合を解消するための訂正候補を求める点では第1の実施の形態の手段12と同じであるが(ステップS4)、図13に示したような結果一覧画面は出力せず、制御を重要度計算手段15に受け渡す。

【０１５５】

重要度計算手段15は、条件判定手段13がリンク不整合として抽出したリンクについて重要度を計算しランキングして出力する（図26のステップS8、S9）。ここで重要度は、(1)検出された箇所の誤り／不適切の種類、(2)検出された箇所の誤り／不適切の確度、(3)検出された箇所を含むページの被リンク数、(4)検出された箇所を含むページに対するユーザからのアクセス実績、(5)検出された箇所を含むページのハイパーテキストにおける階層レベル、のうちの１ファクタもしくは複数のファクタの組み合わせによって計算する。

【０１５６】

出力されるリンク不整合のランキング画面を図27に示す。図13の結果一覧画面と相違するところは、リンク先とリンク元表記が同じリンクをグループ化し、それぞれに不整合の種類、訂正候補に加えて不整合の重要度を付与し、不整合の重要度が高い順に表示している点である。従って、管理者はステップS6における訂正候補の確認や書き換えなどの作業を、不整合の重要度の高い順に実施することが可能である。以下、第1の実施の形態と同様に、管理者に確認ないし修正された訂正候補に基づいて訂正反映手段14によるハイパーテキストデータベース21の各文書の修正が行われる（図26のステップS7）。

【０１５７】

なお、本実施の形態では、候補計算手段12が訂正候補を求めてから重要度計算手段15が重要度を計算してランキングして出力する場合について説明したが、先に重要度計算手段15が重要度を計算してランキングしておき、後で候補計算手段12が訂正候補を求める方法もあり、本実施の形態で述べた方法に限定されない。

【０１５８】

また、本実施の形態では、ステップS6で管理者が出力されたリンク不整合と訂正候補の確認を行ったが、ステップS6を省略してステップS1～S4、S8、S9、S7をすべて自動化して実行する方法もあり、本実施の形態に述べた方法に限定されない。

【０１５９】

また、本実施の形態では、管理者が検査のタイミングを決めて実行する場合について説明したが、あらかじめ収集条件と抽出条件を設定しておき、定期的に自動でステップS1～S4、S8、S9までを実行し、得られた結果をメールなどで通知する方法などもあり、本実施の形態に述べた方法に限定されない。

【０１６０】

なお、本実施の形態のデータ処理装置5は情報収集手段11を備えているが、情報収集手段11によるハイパーテキストデータベース21からのページおよびリンクに関する情報の収集と記憶を別のデータ処理装置で実施する形態も考えられ、そのような形態ではデータ処理装置5の情報収集手段11は省略可能である。また、図27の結果一覧画面を見て管理者自身の手作業によってハイパーテキストデータベース21の不整合箇所を訂正する構成にあっては、訂正反映手段14を省略することができる。この場合、管理者の負担が増大するものの、図27の結果一覧画面の不整合の種類、訂正候補が無くても残りの情報から管理者自身で訂正候補を求めることもできるため、候補計算手段12を省略した構成も採用可能である。

【０１６１】

【発明の第3の実施の形態】

次に、本発明の第3の実施の形態について、図面を参照して詳細に説明する。

【０１６２】

図28を参照すると、本発明の第3の実施の形態は、データ処理装置6が、図25に示された第2の実施の形態におけるデータ処理装置5の構成から、訂正反映手段14を除き、トータルスコア計算手段16を加えた点で第2の実施の形態と異なる。

【０１６３】

トータルスコア計算手段16は、条件判定手段13が出力するリンク不整合と、重要度計算手段15が計算する不整合の重要度とを元に、診断対象のサイトの整合性の

トータルスコアを計算する。ここで、トータルスコアとは、重要度計算手段15が計算する不整合の重要度の合計を使う他に、リンク不整合の数や、総リンク数に対するリンク不整合の数の割合などを使う方法がある。

【０１６４】

以下、本実施の動作を図面を参照して詳細に説明する。

【０１６５】

図29のステップS1～S4、S8で示される本実施の形態における情報収集手段11、候補計算手段12、条件判定手段13、重要度計算手段15の動作は、第2の実施の形態の各手段11、12、13、15の動作と同一のため、説明は省略する。

【０１６６】

第2の実施の形態では、リンク不整合を検知した後、訂正候補に従ってハイパーテキストデータベース21に訂正を反映していた。本実施の形態では、リンクの不整合を検知した後、重要度計算手段15が求めた重要度を基に、トータルスコア計算手段16が診断対象サイト全体のトータルスコアを計算して出力する（図29のステップS10）。

【０１６７】

このトータルスコア計算を定期的に行い、時間変化を図30のように出力することによって、サイト品質の改善経過を知ることができる。図30では、時間とともにトータルスコアの上昇が飽和してきており、サイト品質の改善作業が収束に近づいていることがわかる。

【０１６８】

また、このトータルスコア計算を定期的に行い、トータルスコアあるいはリンク不整合として検出された箇所の重要度が閾値を超えるなど、あらかじめ定めた条件を満たした場合にアラートを通知することによって、サイト品質が低下した場合には、サイト管理者は警告を受け取ることができる。

【０１６９】

また、トータルスコア計算を複数の異なるサイトA～Mについて行い、図31のようにランキングして出力することによって、サイト品質を定量的に比較することができる。図31では、サイトAの品質はサイトEの2倍程度優れていることがわかる

。

【０１７０】

次に、本実施の形態の効果について説明する。

【０１７１】

本実施の形態では、リンク不整合の検知数や重要度を基に診断対象とするサイト品質のトータルスコアを計算する。そのため、サイト品質の改善経過を把握したり、異なるサイト間の品質を定量的に比較することができる。

【０１７２】

なお、本実施の形態のデータ処理装置6は情報収集手段11を備えているが、情報収集手段11によるハイパーテキストデータベース21からのページおよびリンクに関する情報の収集と記憶を別のデータ処理装置で実施する形態も考えられ、そのような形態ではデータ処理装置6の情報収集手段11は省略可能である。また、本実施の形態の説明においては、検出された不整合箇所のハイパーテキストデータベース21への反映（訂正）については触れなかったが、反映を行うようにしても、行わないようにしても良い。反映する場合、図27の結果一覧画面を見て管理者自身の手作業によってハイパーテキストデータベース21の不整合箇所を訂正するようにしても良いし、第2の実施の形態と同様な訂正反映手段14を設ける構成も考えられる。さらに、管理者の負担が増大するものの、図27の結果一覧画面の不整合の種類、訂正候補が無くても残りの情報から管理者自身で訂正候補を求めることもできるため、候補計算手段12を省略した構成も採用可能である。

【０１７３】

【発明の第4の実施の形態】

次に、本発明の第4の実施の形態について、図面を参照して詳細に説明する。

【０１７４】

図32を参照すると本発明に係る第4の実施の形態は、本発明の第1の実施の形態と同様に、入力手段501、データ処理装置502、出力手段503、記憶装置504を備える。さらに、第1の実施の形態のキーワード抽出装置を実現するためのハイパーテキスト検査用プログラム500を備える。

【０１７５】

入力手段501は、マウス、キーボード等、操作者からの指示を入力するための装置である。また、出力手段503は、表示画面、プリンタ等のデータ処理装置502による処理結果を出力する装置である。

【０１７６】

ハイパーテキスト検査用プログラム500は、データ処理装置502に読み込まれ、データ処理装置502の動作を制御し、記憶装置504に入力メモリ505とワークメモリ506を生成すると共に、データ処理装置502上に第1の実施形態における図1の情報収集手段11、候補計算手段12、条件判定手段13および訂正反映手段14を実現する。データ処理装置502は、ハイパーテキスト検査装置を実現するためのプログラムの制御により第1の実施形態と同一の処理を実行する。

【０１７７】

図1におけるデータ処理装置1と図32におけるデータ処理装置502が対応し、図1における記憶装置2と図32における記憶装置504が対応する。ただし、処理対象となるハイパーテキストデータベース21は、記憶装置504に格納されたデータを利用する他に、データ処理装置502によって外部にあるデータベースにネットワーク（例えばインターネット）を介してアクセスして取得する形態であってもよい。

【０１７８】

【発明の第5の実施の形態】

次に、本発明の第5の実施の形態について、図面を参照して詳細に説明する。

【０１７９】

第5の実施の形態は、第4の実施の形態と同様に、図32の構成を用いる。ハイパーテキスト検査用プログラム500は、データ処理装置502に読み込まれ、データ処理装置502の動作を制御し、記憶装置504に入力メモリ505とワークメモリ506を生成すると共に、データ処理装置502上に第2の実施形態における図25の情報収集手段11、候補計算手段12、条件判定手段13、訂正反映手段14および重要度計算手段15を実現する。データ処理装置502は、ハイパーテキスト検査装置を実現するためのプログラムの制御により第2の実施形態と同一の処理を実行する。

【０１８０】

図25におけるデータ処理装置5と図32におけるデータ処理装置502が対応し、図25

における記憶装置2と図32における記憶装置504が対応する。ただし、処理対象となるハイパーテキストデータベース21は、記憶装置504に格納されたデータを利用する他に、データ処理装置502によって外部にあるデータベースにネットワーク（例えばインターネット）を介してアクセスして取得する形態であってもよい。

【０１８１】

【発明の第6の実施の形態】

次に、本発明の第6の実施の形態について、図面を参照して詳細に説明する。

【０１８２】

第6の実施の形態は、第4の実施の形態と同様に、図32の構成を用いる。ハイパーテキスト検査用プログラム500は、データ処理装置502に読み込まれ、データ処理装置502の動作を制御し、記憶装置504に入力メモリ505とワークメモリ506を生成すると共に、データ処理装置502上に第3の実施形態における図28の情報収集手段11、候補計算手段12、条件判定手段13、重要度計算手段15およびトータルスコア計算手段16を実現する。データ処理装置502は、ハイパーテキスト検査装置を実現するためのプログラムの制御により第3の実施形態と同一の処理を実行する。

【０１８３】

図28におけるデータ処理装置6と図32におけるデータ処理装置502が対応し、図28における記憶装置2と図32における記憶装置504が対応する。ただし、処理対象となるハイパーテキストデータベース21は、記憶装置504に格納されたデータを利用する他に、データ処理装置502によって外部にあるデータベースにネットワーク（例えばインターネット）を介してアクセスして取得する形態であってもよい。

【０１８４】

【発明の効果】

以上説明したように本発明によれば以下のような効果が得られる。

【０１８５】

第1の効果は、各種の論理的不整合を検知することができる。その理由は、例えば、ハイパーテキストデータベースからリンク情報を抽出し、リンク情報の各項

目毎にリンクをグループすることにより、グループから外れた特異なリンクをリンク不整合として検知するため、(1)リンクの張り間違い、(2)期限切れ情報へのリンク、(3)リンク元表記の不統一、(4)リンク元表記のスタイルの不統一といった論理的不整合を検知することができるからである。また、リンク情報の収集を定期的に繰り返し実行し、リンク情報の時系列変化に着目して各種リンク不整合を検知する方法によって、(2)期限切れ情報へのリンクといった論理的不整合を検知することができるからである。更に、リンク元表記のないリンクの検出により、(5)幽霊リンクといった論理的不整合を検知することができるからである。また更に、ループを形成するリンクの系列であって、かつ、該リンクの系列に対応するリンク元表記がすべて同一トピックに関わるものを検出することにより、(6)ループリンクといった論理的不整合を検知することができるからである。

【０１８６】

第2の効果は、リンク不整合の訂正方法を自動で決定できるために、管理者は不整合をどのように修正すべきかを検討する必要がないことである。その理由は、例えば、特異なリンクのリンク情報を、グループと同一のリンク情報にするように訂正候補を自動計算するからである。

【０１８７】

第3の効果は、管理者によるチェックの効率が大幅に向上することである。その理由は、リンク不整合をグループ化してまとめて表示するため、管理者は一部のリンクを確認すれば残りのリンクも同様に不整合か否かを判定できるからである。

【０１８８】

第4の効果は、ページ単位での修正項目を把握したり、重要なページに対する不整合を重点的に調べたり、リンク元表記として使っている表現の妥当性などを調べることができることである。その理由は、リンク不整合を（１）リンク元表記、（２）リンク元ページの識別情報、（３）リンク先ページの識別情報、という３項目のいずれかをソートキーとして一覧表示するからである。

【０１８９】

第5の効果は、サイト品質の改善経過を把握したり、異なるサイト間の品質を定

量的に比較することができることである。その理由は、検知されたリンク不整合の数や重要度を基に診断対象とするサイトのトータルスコアを計算するからである。

【図面の簡単な説明】

【図1】

本発明の第1の実施の形態の構成を示すブロック図である。

【図2】

ハイパーテキストにおけるリンクの指定方法とブラウザ上での表示例を示す図である。

【図3】

間違いリンクによる論理的不整合の例を示す図である。

【図4】

期限切れリンクによる論理的不整合の例を示す図である。

【図5】

リンク元表記の不統一による論理的不整合の例を示す図である。

【図6】

リンク元表記のスタイルの不統一による論理的不整合の例を示す図である。

【図7】

幽霊リンクによる論理的不整合の例を示す図である。

【図8】

ループリンクによる論理的不整合の例を示す図である。

【図9】

情報記憶部に格納されるリンク情報の例を示す図である。

【図10】

本発明の第1の実施の形態の動作を示す流れ図である。

【図11】

本発明の第1の実施の形態における文書収集条件の設定画面例を示す図である。

【図12】

本発明の第1の実施の形態におけるリンク不整合の抽出条件の設定画面例を示す

図である。

【図13】

本発明の第1の実施の形態における不整合抽出結果の例を示す図である。

【図14】

本発明の第1の実施の形態における間違いリンク抽出の動作を示す流れ図である。

【図15】

本発明の第1の実施の形態における間違いリンク抽出時に抽出されるリンク情報の例を示す図である。

【図16】

本発明の第1の実施の形態における期限切れリンク抽出の動作を示す流れ図である。

【図17】

本発明の第1の実施の形態におけるリンク元表記の不統一抽出の動作を示す流れ図である。

【図18】

本発明の第1の実施の形態におけるリンク元表記の不統一抽出時に抽出されるリンク情報の例を示す図である。

【図19】

本発明の第1の実施の形態におけるリンク元表記のスタイルの不統一抽出の動作を示す流れ図である。

【図20】

本発明の第1の実施の形態におけるリンク元表記のスタイルの不統一抽出時に抽出されるリンク情報の例を示す図である。

【図21】

本発明の第1の実施の形態における幽霊リンク抽出の動作を示す流れ図である。

【図22】

本発明の第1の実施の形態におけるループリンク抽出の動作を示す流れ図である。

【図23】

本発明の第1の実施の形態におけるリンク情報の時間変化抽出の動作を示す流れ図である。

【図24】

本発明の第1の実施の形態におけるリンク情報の時間変化抽出時に抽出されるリンク情報の例を示す図である。

【図25】

本発明の第2の実施の形態の構成を示すブロック図である。

【図26】

本発明の第2の実施の形態の動作を示す流れ図である。

【図27】

本発明の第2の実施の形態における不整合抽出結果の例を示す図である。

【図28】

本発明の第3の実施の形態の構成を示すブロック図である。

【図29】

本発明の第3の実施の形態の動作を示す流れ図である。

【図30】

本発明の第3の実施の形態におけるトータルスコアの時間変化を出力する画面例を示す図である。

【図31】

本発明の第3の実施の形態におけるトータルスコアによるサイトのランキング画面例を示す図である。

【図32】

本発明の第4、第5および第6の実施の形態の構成を示すブロック図である。

【符号の説明】

1、5、6、502 データ処理装置

2、504 記憶装置

3、501 入力手段

4、503 出力手段

11 情報収集手段

12 候補計算手段

13 条件判定手段

14 訂正反映手段

15 重要度計算手段

16 トータルスコア計算手段

21 ハイパーテキストデータベース

22 情報記憶部

500 ハイパーテキスト検査用プログラム

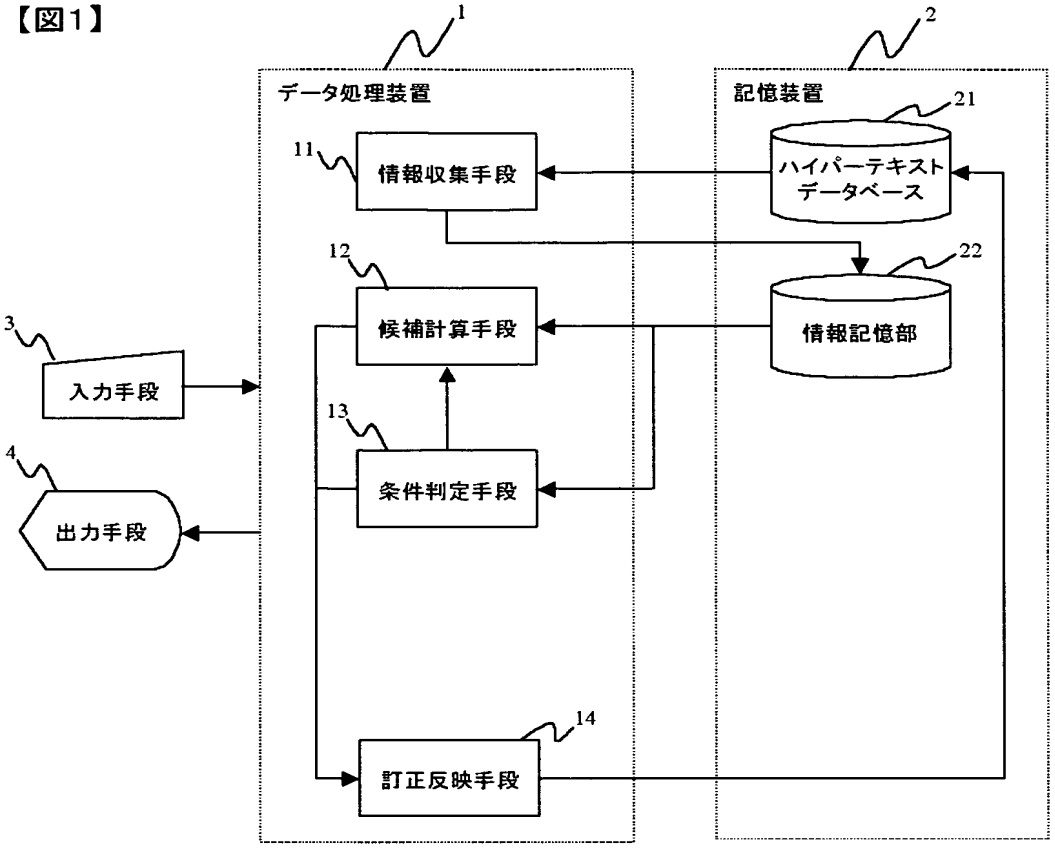
505 入力メモリ

506 ワークメモリ

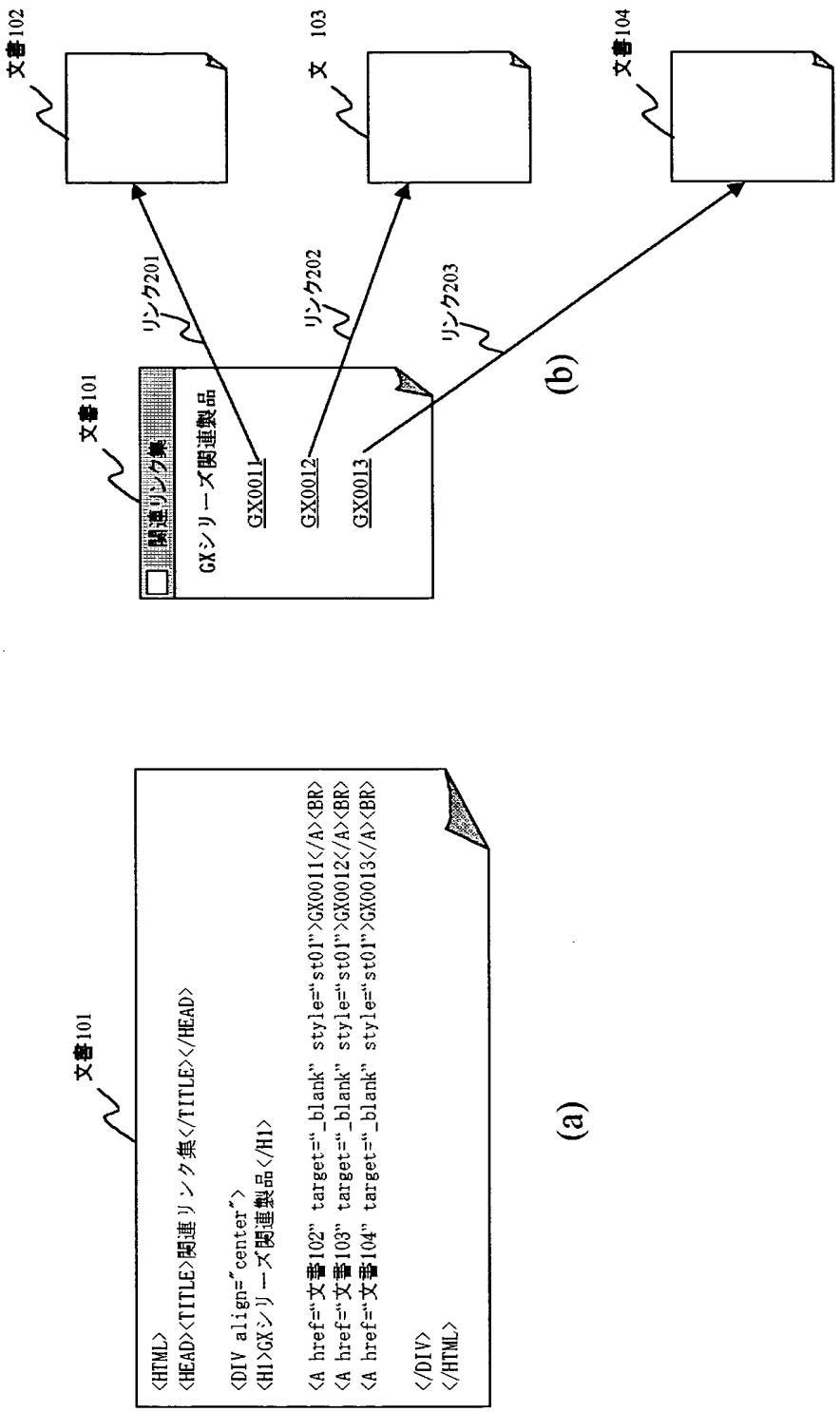
【書類名】　図面

【図１】

【図1】

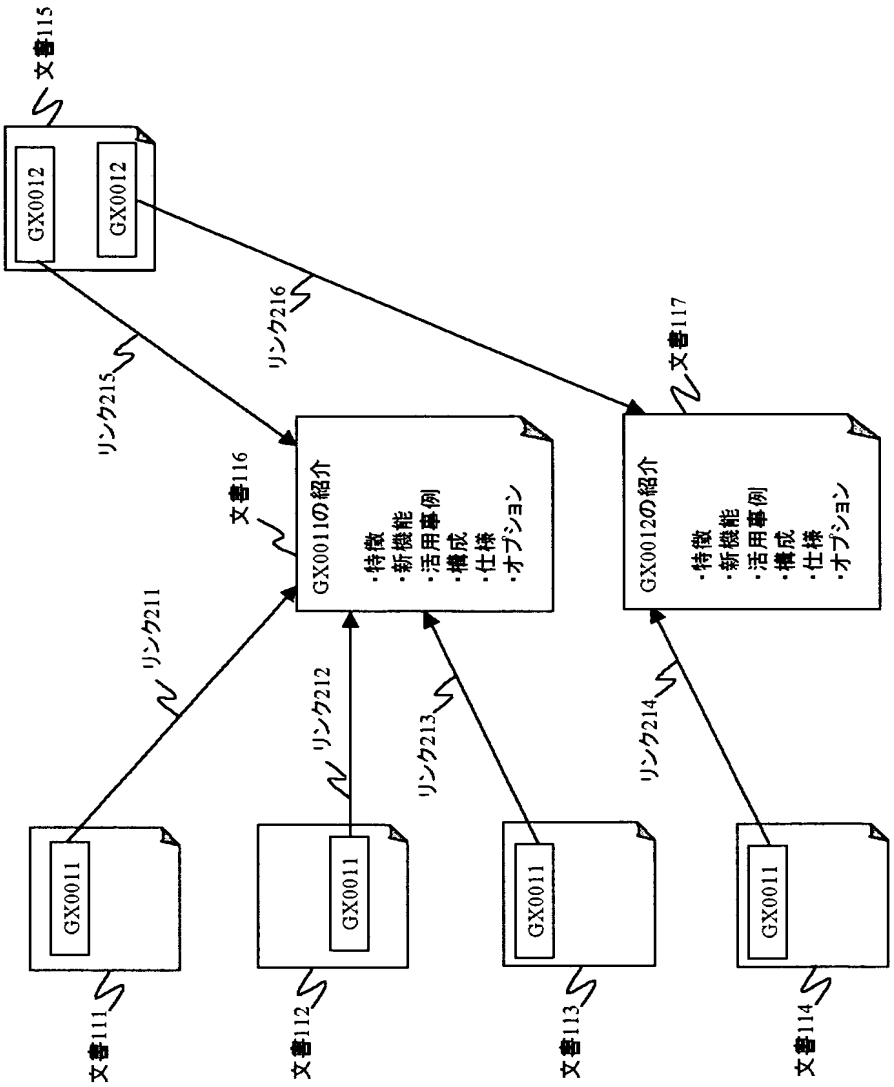

1
2
データ処理装置
記憶装置
11
情報収集手段
21
ハイパーテキスト
データベース
12
候補計算手段
22
情報記憶部
3
入力手段
13
条件判定手段
4
出力手段
14
訂正反映手段

【図2】


文書102
文書103
文書104
文書101
リンク201
リンク202
リンク203
関連リンク集
GXシリーズ関連製品
GX0011
GX0012
GX0013
(b)
文書101
<HTML>
<HEAD><TITLE>関連リンク集</TITLE></HEAD>
<DIV align="center">
<H1>GXシリーズ関連製品</H1>
<A href="文書102" target="_blank" style="st01">GX0011</A>
<A href="文書103" target="_blank" style="st01">GX0012</A>
<A href="文書104" target="_blank" style="st01">GX0013</A>
</DIV>
</HTML>
(a)

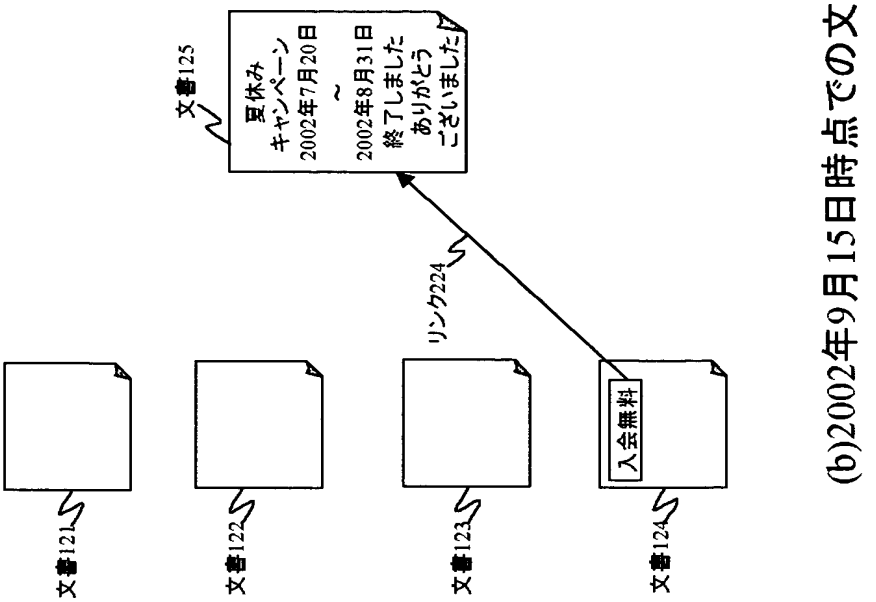
【図３】

【図3】

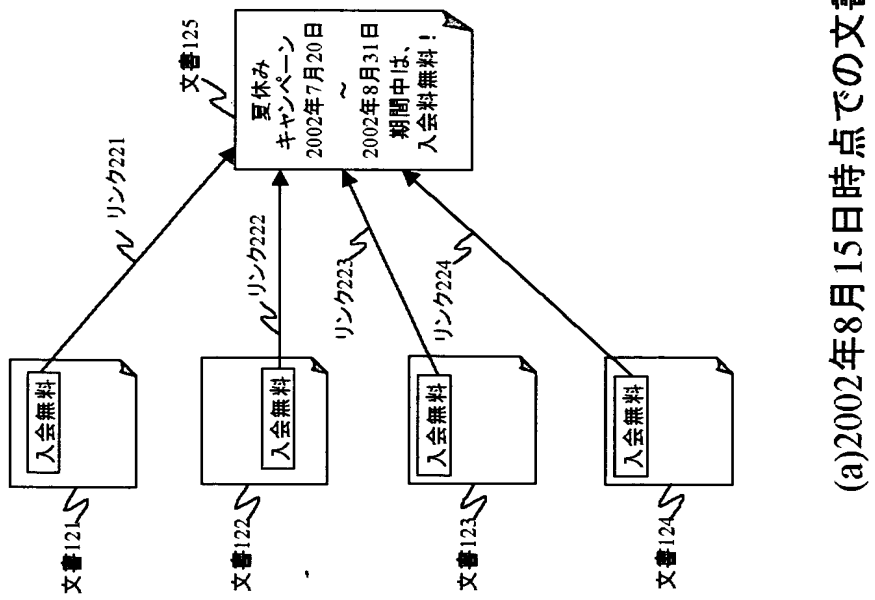

文書115
GX0012
GX0012
リンク215
リンク216
文書116
GX0011の紹介
・特徴
・新機能
・活用事例
・構成
・仕様
・オプション
文書117
GX0012の紹介
・特徴
・新機能
・活用事例
・構成
・仕様
・オプション
リンク211
リンク212
リンク213
リンク214
GX0011
GX0011
GX0011
GX0011
文書111
文書112
文書113
文書114

【図４】

【図4】

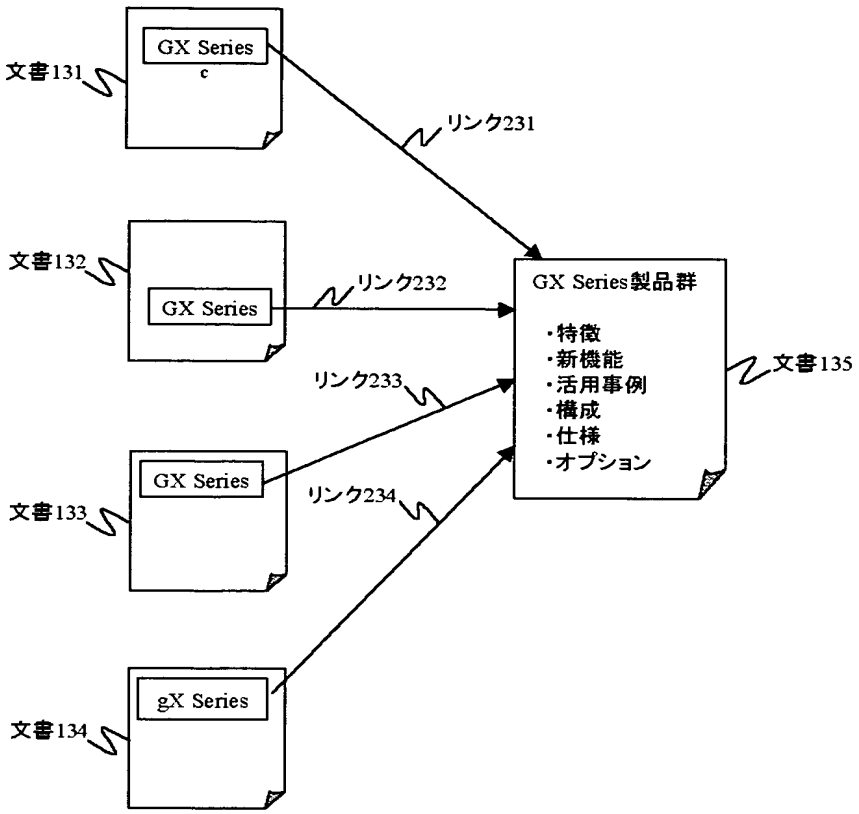


(b)2002年9月15日時点での文書群

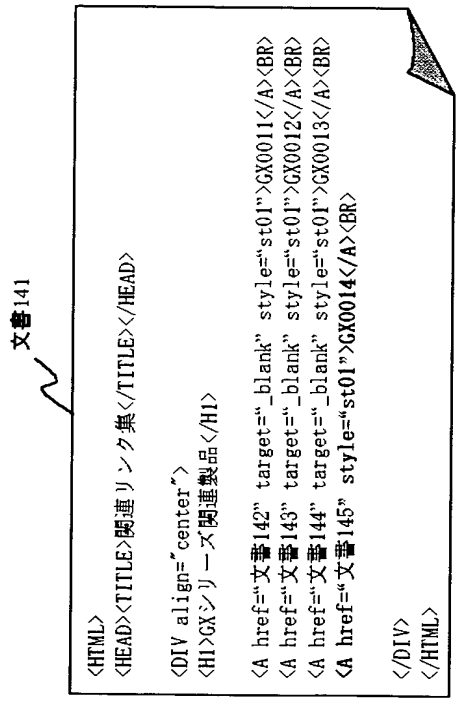


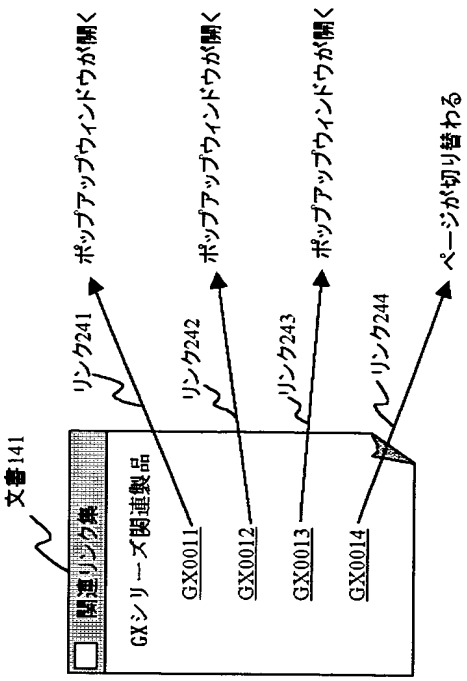
(a)2002年8月15日時点での文書群

【図5】


GX Series
c
文書131
リンク231
文書132
GX Series
リンク232
GX Series製品群
・特徴
・新機能
・活用事例
・構成
・仕様
・オプション
文書135
リンク233
GX Series
文書133
リンク234
gX Series
文書134

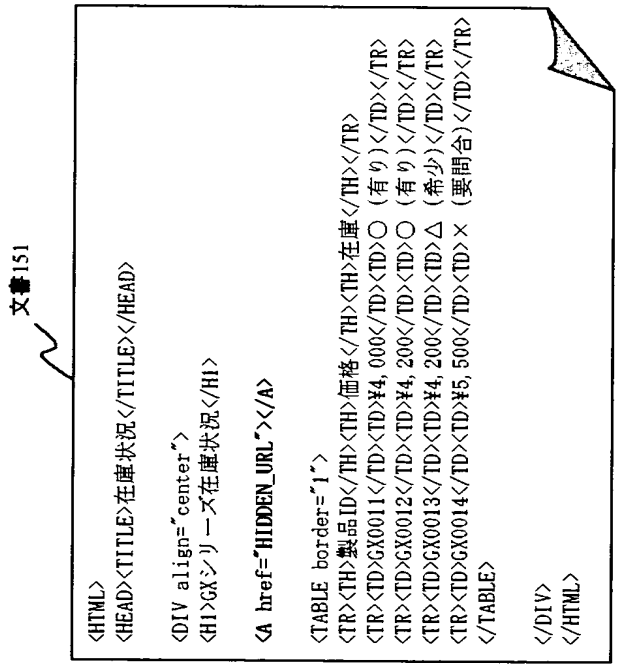
【図6】


文書141
<HTML>
<HEAD><TITLE>関連リンク集</TITLE></HEAD>
<DIV align="center">
<H1>GXシリーズ関連製品</H1>
<A href="文書142" target="_blank" style="st01">GX0011</A>
<A href="文書143" target="_blank" style="st01">GX0012</A>
<A href="文書144" target="_blank" style="st01">GX0013</A>
<A href="文書145" style="st01">GX0014</A>
</DIV>
</HTML>
(a)


文書141
関連リンク集
GXシリーズ関連製品
GX0011
GX0012
GX0013
GX0014
リンク241
リンク242
リンク243
リンク244
ポップアップウィンドウが開く
ポップアップウィンドウが開く
ポップアップウィンドウが開く
ページが切り替わる
(b)

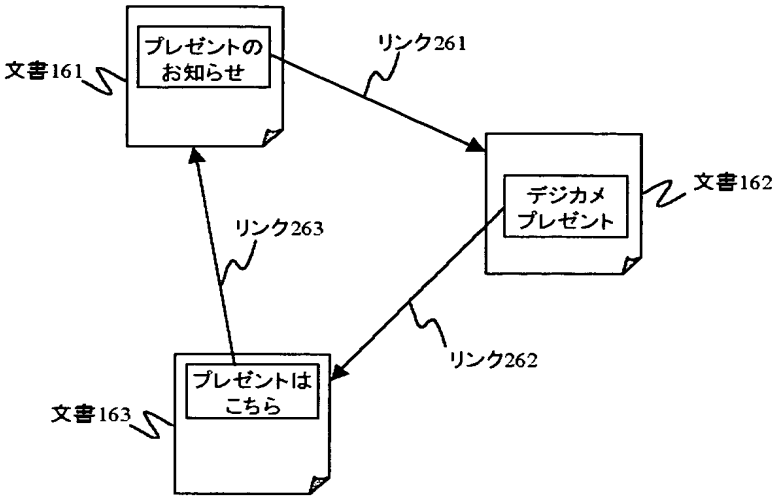
【図７】

【図7】


文書151
<HTML>
<HEAD><TITLE>在庫状況</TITLE></HEAD>
<DIV align="center">
<H1>GXシリーズ在庫状況</H1>
<A href="HIDDEN_URL"></A>
<TABLE border="1">
<TR><TH>製品ID</TH><TH>価格</TH><TH>在庫</TH></TR>
<TR><TD>GX0011</TD><TD>¥4,000</TD><TD>○（有り）</TD></TR>
<TR><TD>GX0012</TD><TD>¥4,200</TD><TD>○（有り）</TD></TR>
<TR><TD>GX0013</TD><TD>¥4,200</TD><TD>△（希少）</TD></TR>
<TR><TD>GX0014</TD><TD>¥5,500</TD><TD>×（要問合）</TD></TR>
</TABLE>
</DIV>
</HTML>
(a)
文書151
在庫状況
GXシリーズ在庫状況
製品ID | 価格 | 在庫
GX0011 | ¥4,000 | ○（有り）
GX0012 | ¥4,200 | ○（有り）
GX0013 | ¥4,200 | △（希少）
GX0014 | ¥5,500 | ×（要問合）
(b)

【図８】

【図8】



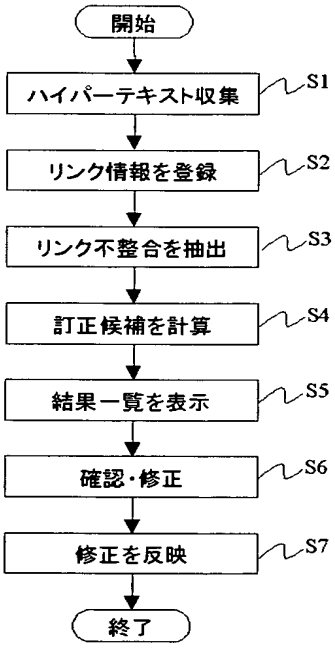
【図９】

【図9】

リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
リンク201	文書101	文書102	GX0011	_blank	st01
リンク202	文書101	文書103	GX0012	_blank	st01
リンク203	文書101	文書104	GX0013	_blank	st01

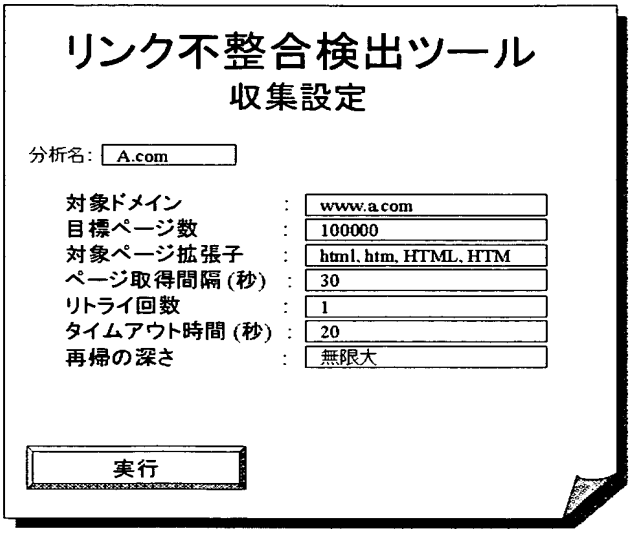
【図１０】

【図10】



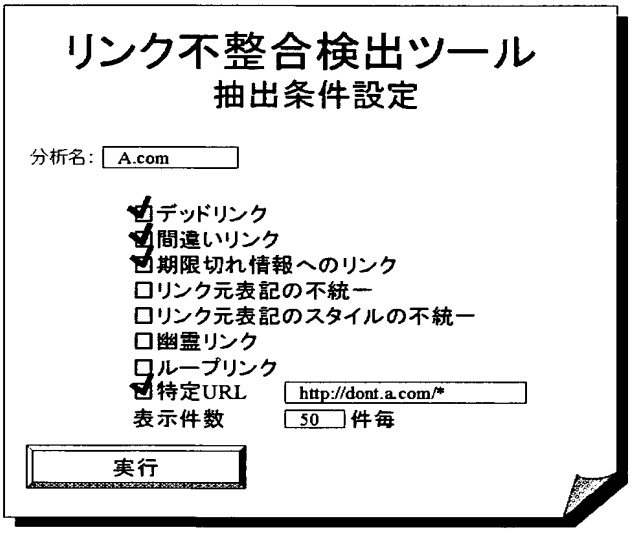
【図１１】

【図11】



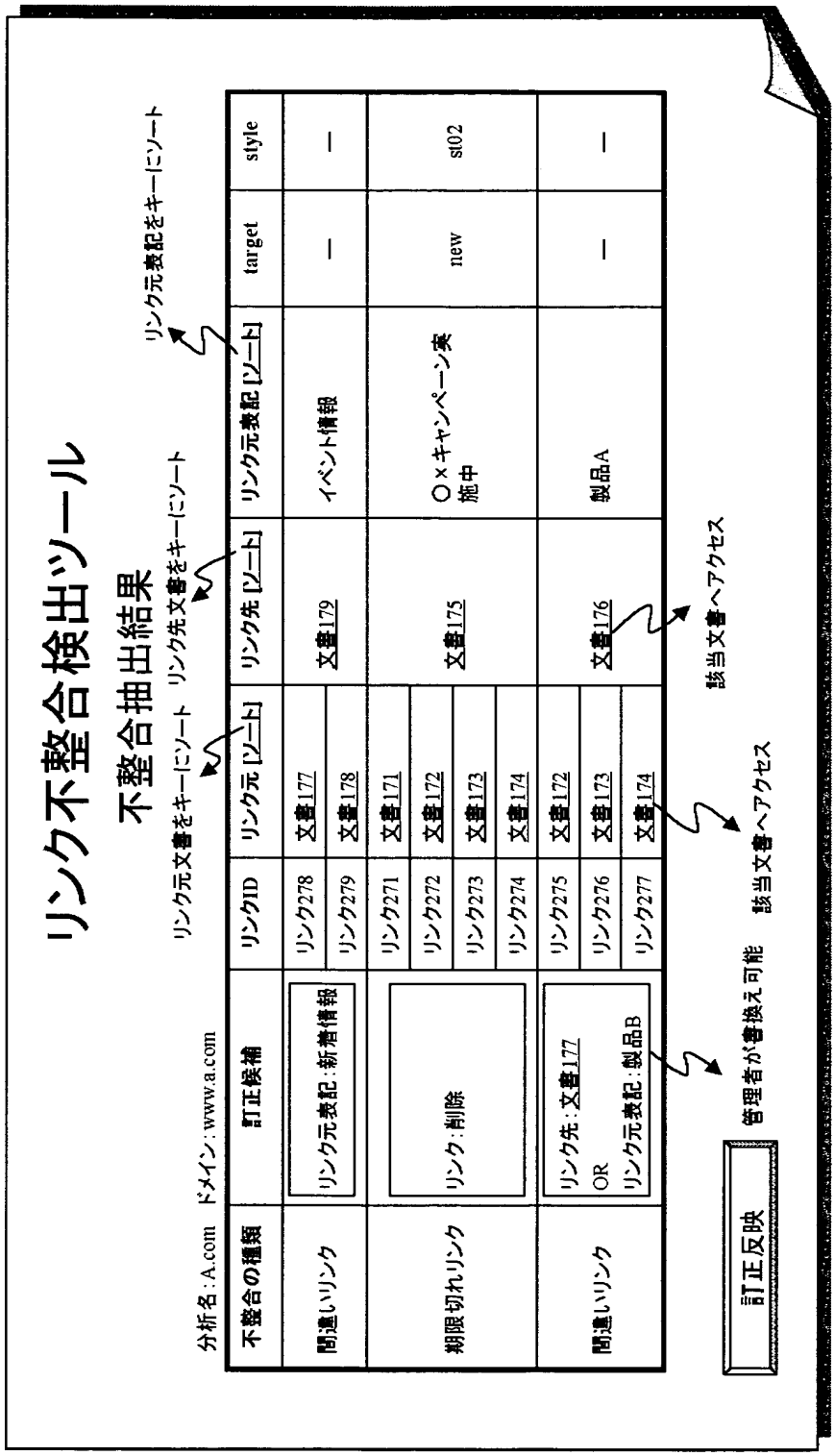
【図１２】

【図12】



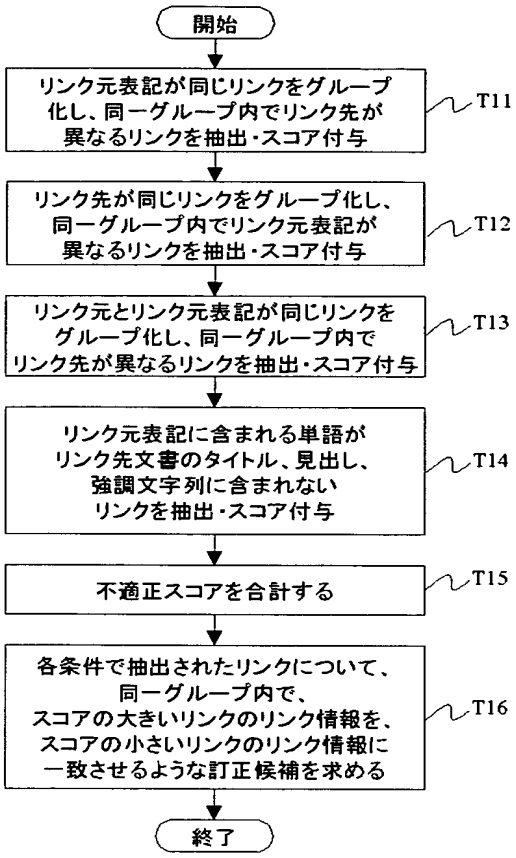
【図１３】

【図13】


リンク不整合検出ツール
不整合抽出結果
分析名：A.com　ドメイン：www.a.com
リンク元文書をキーにソート
リンク先文書をキーにソート
リンク元表記をキーにソート
不整合の種類
訂正候補
リンクID
リンク元 [ソート]
リンク先 [ソート]
リンク元表記 [ソート]
target
style
間違いリンク
リンク元表記：新着情報
リンク278
リンク279
文書177
文書178
文書179
イベント情報
—
—
期限切れリンク
リンク：削除
リンク271
リンク272
リンク273
リンク274
文書171
文書172
文書173
文書174
文書175
○×キャンペーン実施中
new
st02
間違いリンク
リンク先：文書177
OR
リンク元表記：製品B
リンク275
リンク276
リンク277
文書172
文書173
文書174
文書176
製品A
—
—
訂正反映
管理者が書換え可能
該当文書へアクセス
該当文書へアクセス

【図１４】

【図14】


開始
リンク元表記が同じリンクをグループ化し、同一グループ内でリンク先が異なるリンクを抽出・スコア付与
T11
リンク先が同じリンクをグループ化し、同一グループ内でリンク元表記が異なるリンクを抽出・スコア付与
T12
リンク元とリンク元表記が同じリンクをグループ化し、同一グループ内でリンク先が異なるリンクを抽出・スコア付与
T13
リンク元表記に含まれる単語がリンク先文書のタイトル、見出し、強調文字列に含まれないリンクを抽出・スコア付与
T14
不適正スコアを合計する
T15
各条件で抽出されたリンクについて、同一グループ内で、スコアの大きいリンクのリンク情報を、スコアの小さいリンクのリンク情報に一致させるような訂正候補を求める
T16
終了

【図１５】

【図15】

(a)

不適正スコア	リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
1/12	リンク211	文書111	文書116	GX0011	—	—
1/12	リンク212	文書112			—	—
1/12	リンク213	文書113			—	—
3/4	リンク214	文書114	文書117		—	—
1/2	リンク215	文書115	文書116	GX0012	—	—
1/2	リンク216	文書115	文書117		—	—

(b)

不適正スコア	リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
1/12	リンク211	文書111	文書116	GX0011	—	—
1/12	リンク212	文書112			—	—
1/12	リンク213	文書113			—	—
3/4	リンク215	文書115		GX0012	—	—
1/2	リンク214	文書114	文書117	GX0011	—	—
1/2	リンク216	文書115		GX0012	—	—

(c)

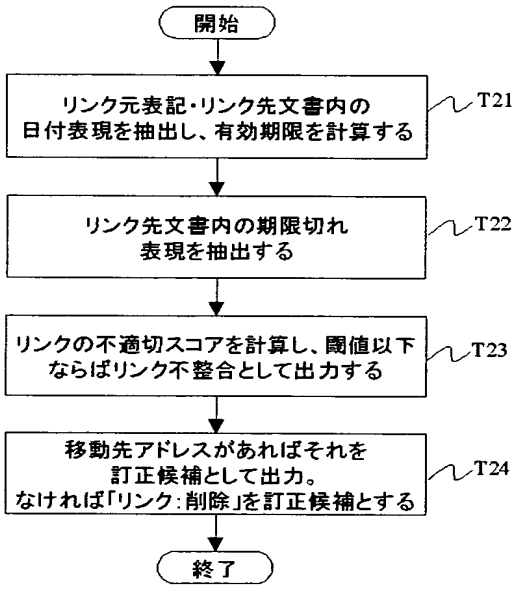
不適正スコア	リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
1/2	リンク215	文書115	文書116	GX0012	—	—
1/2	リンク216		文書117		—	—

(d)

不適正スコア	リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
1	リンク214	文書114	文書117	GX0011	—	—
1	リンク215	文書115	文書116	GX0012	—	—

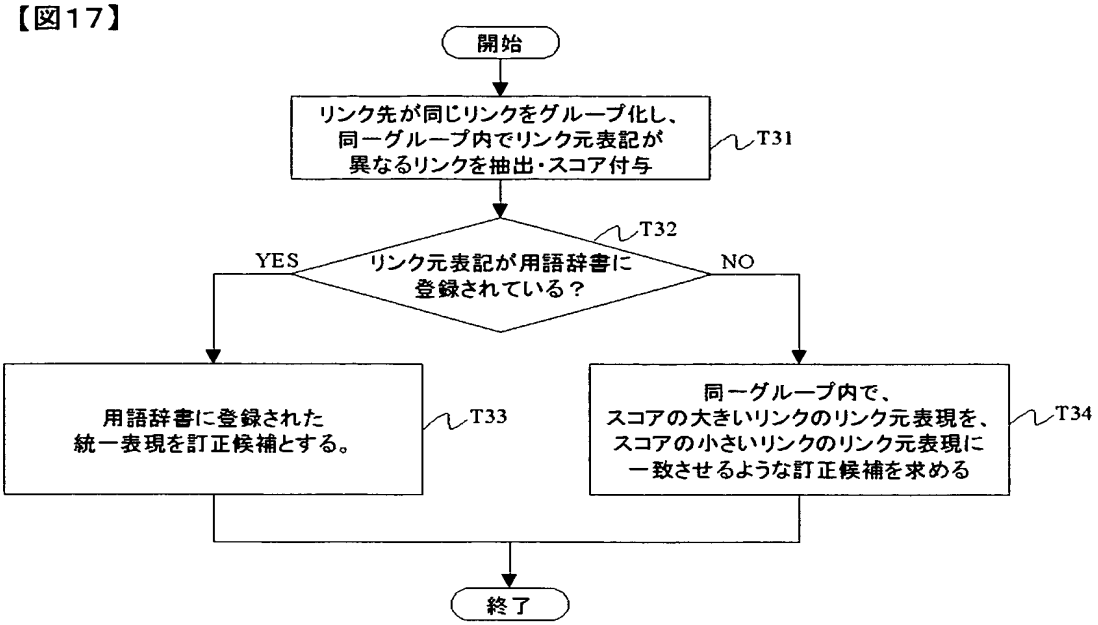
【図１６】

【図16】



【図１７】

【図17】



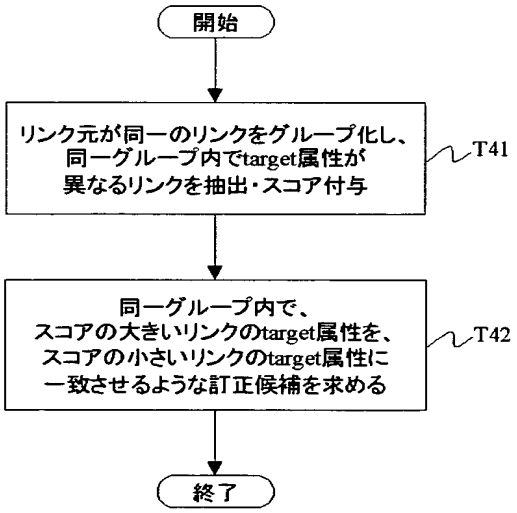
【図１８】

【図18】

不適正スコア	リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
1/12	リンク231	文書131	文書135	GX Series	—	—
1/12	リンク232	文書132			—	—
1/12	リンク233	文書133			—	—
3/4	リンク234	文書134		gX Series	—	—

【図１９】

【図19】



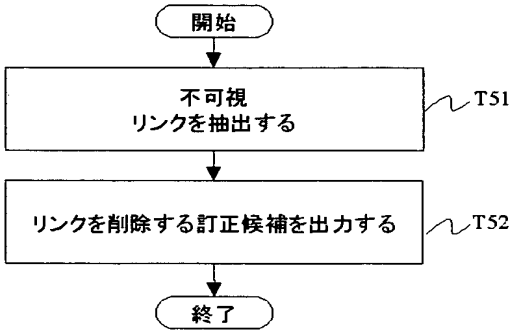
【図２０】

【図20】

不適正スコア	リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
1/12	リンク241	文書141	文書142	GX0011	_blank	st01
1/12	リンク242		文書143	GX0012		st01
1/12	リンク243		文書144	GX0013		st01
3/4	リンク244		文書145	GX0014	—	st01

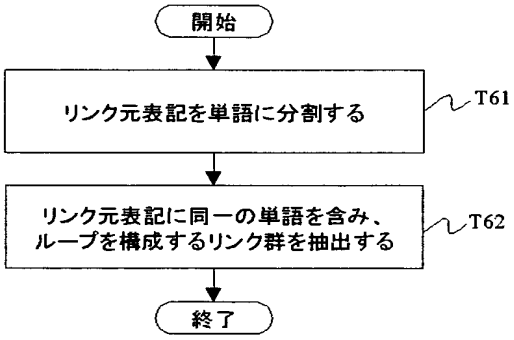
【図２１】

【図21】



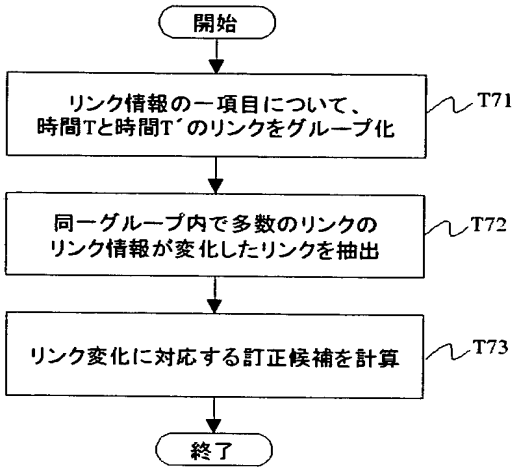
【図２２】

【図22】



【図２３】

【図23】



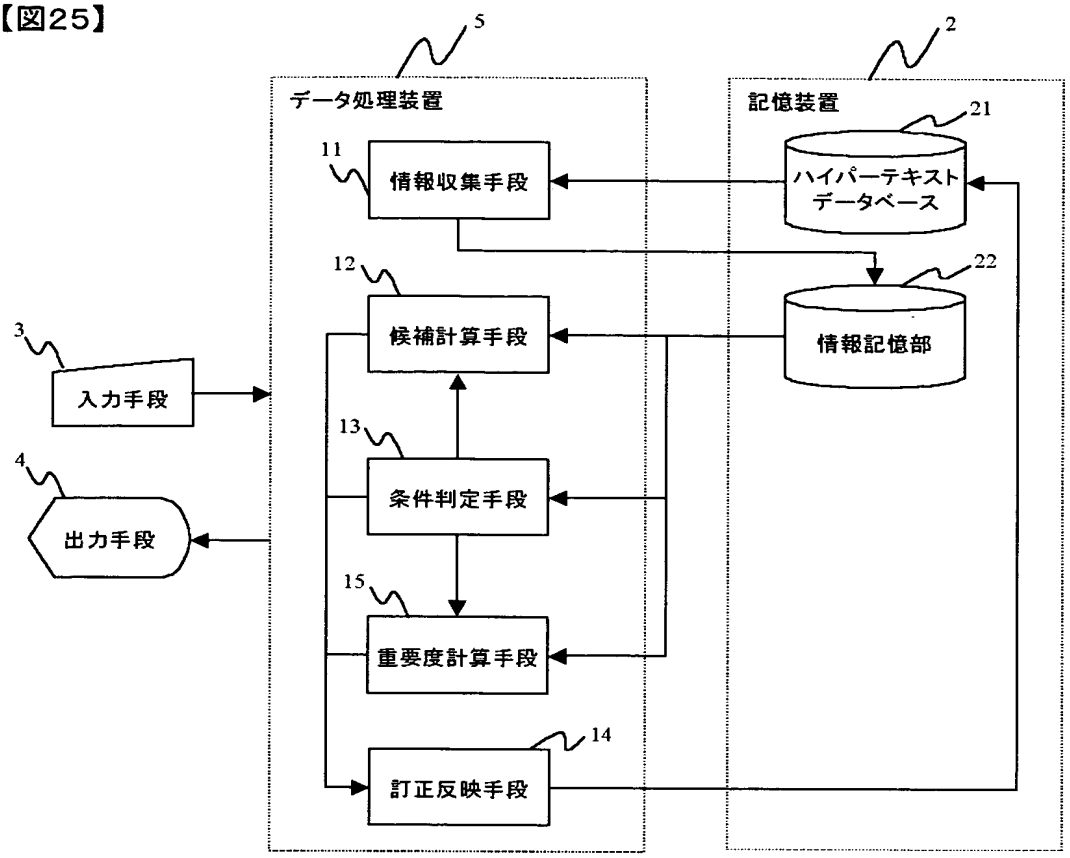
【図２４】

【図24】

時間	リンクID	リンク元アドレス	リンク先アドレス	リンク元表記	target属性	style属性
2002年8月15日	リンク221	文書121	文書125	入会無料	—	—
	リンク222	文書122		入会無料	—	—
	リンク223	文書123		入会無料	—	—
	リンク224	文書124		入会無料	—	—
2002年9月15日	リンク224	文書124		入会無料	—	—

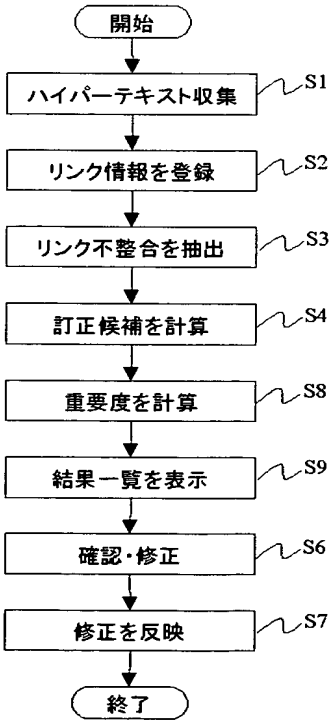
【図２５】

【図25】



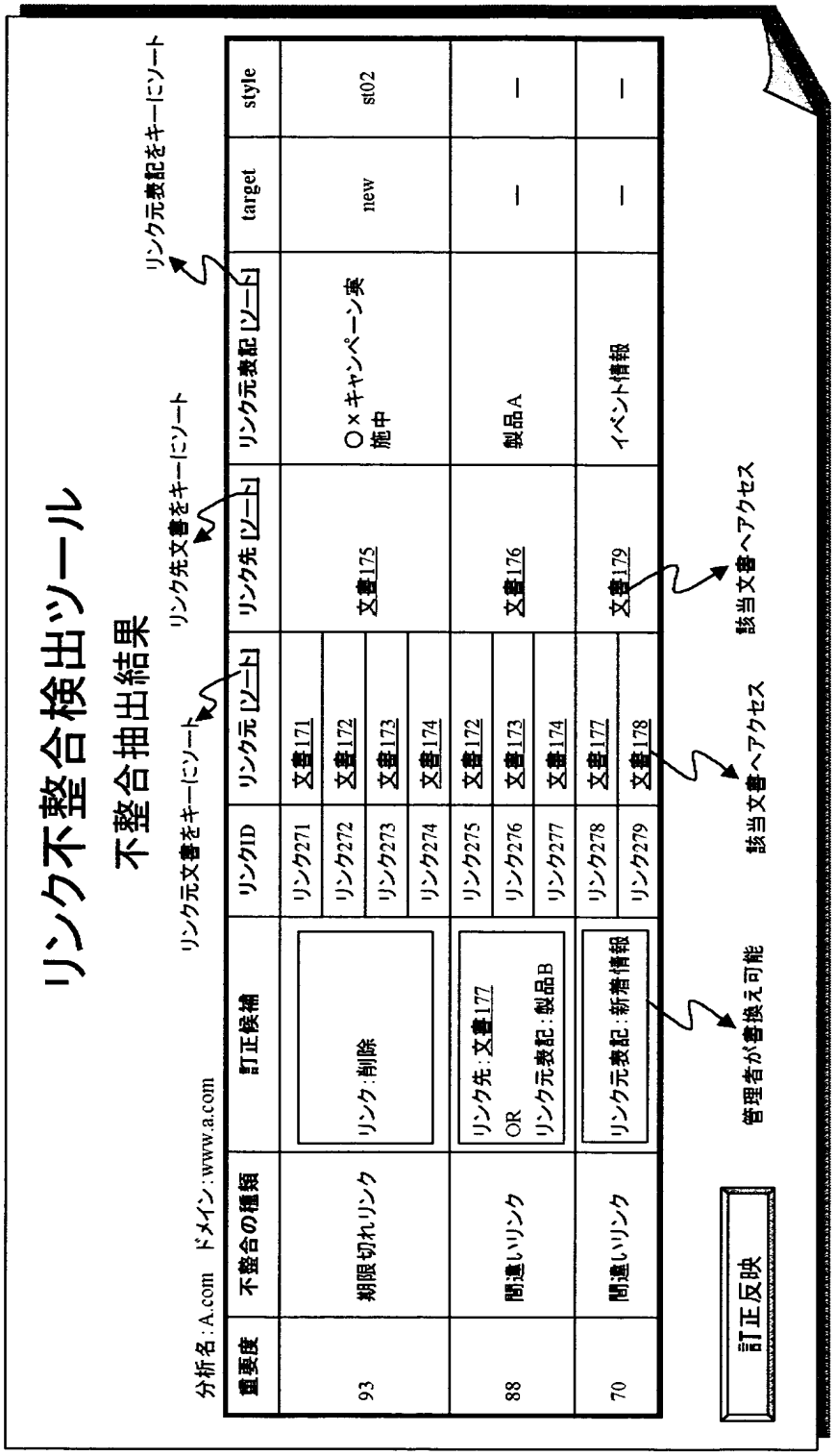
【図２６】

【図26】



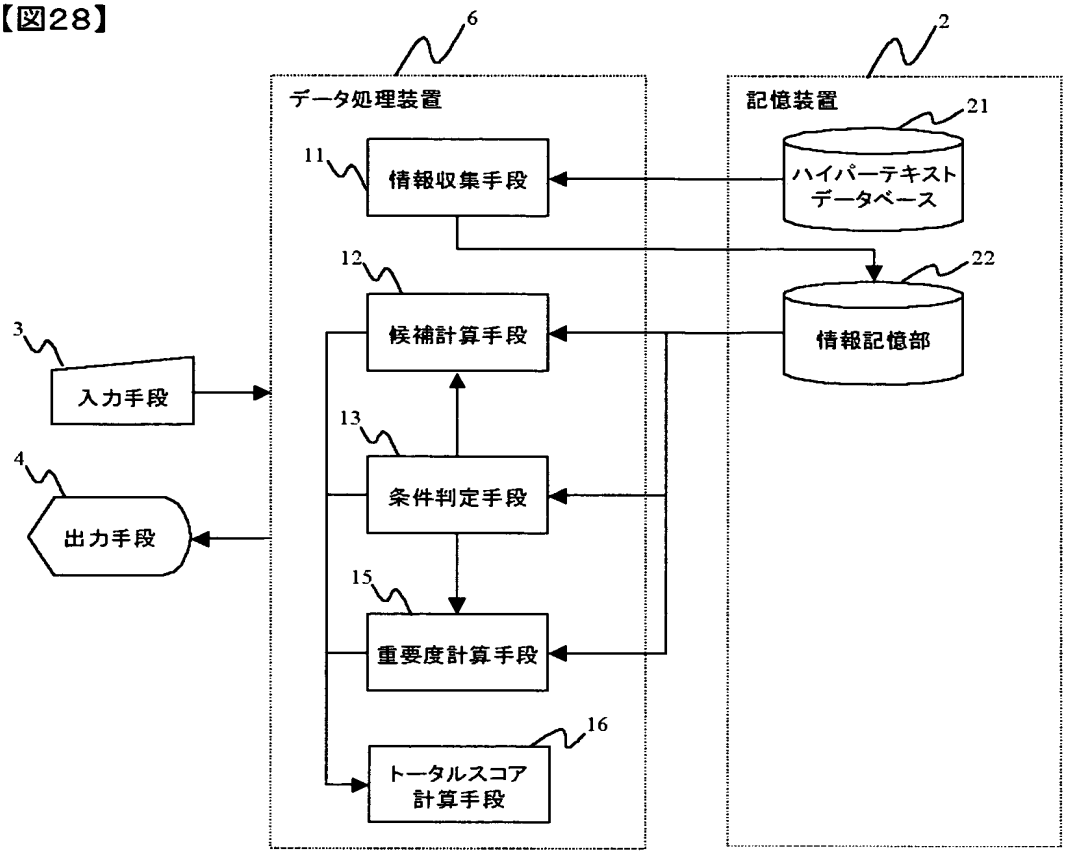
【図２７】

【図27】


リンク不整合検出ツール
不整合抽出結果
分析名：A.com　ドメイン：www.a.com
リンク元文書をキーにソート
リンク先文書をキーにソート
リンク元表記をキーにソート
重要度
不整合の種類
訂正候補
リンクID
リンク元 [ソート]
リンク先 [ソート]
リンク元表記 [ソート]
target
style
93
期限切れリンク
リンク：削除
リンク271
リンク272
リンク273
リンク274
文書171
文書172
文書173
文書174
文書175
○×キャンペーン実施中
new
st02
88
間違いリンク
リンク先：文書177
OR
リンク元表記：製品B
リンク275
リンク276
リンク277
文書172
文書173
文書174
文書176
製品A
—
—
70
間違いリンク
リンク元表記：新着情報
リンク278
リンク279
文書177
文書178
文書179
イベント情報
—
—
訂正反映
管理者が書換え可能
該当文書へアクセス
該当文書へアクセス

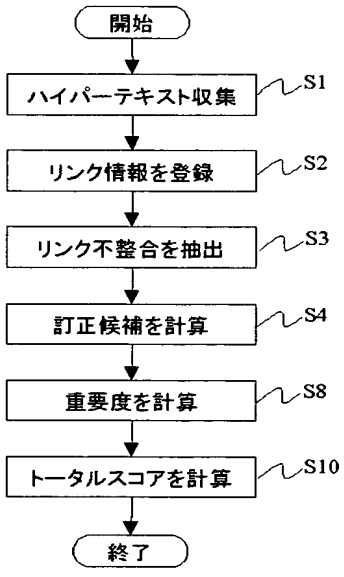
【図２８】

【図28】



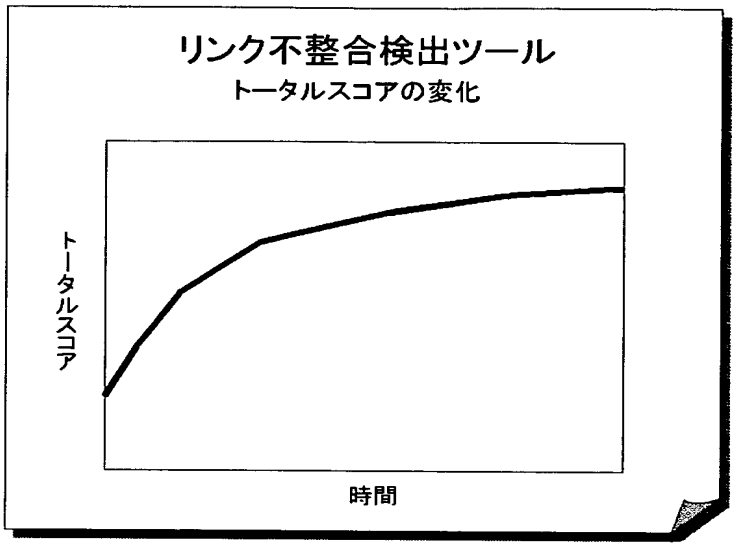
【図２９】

【図29】



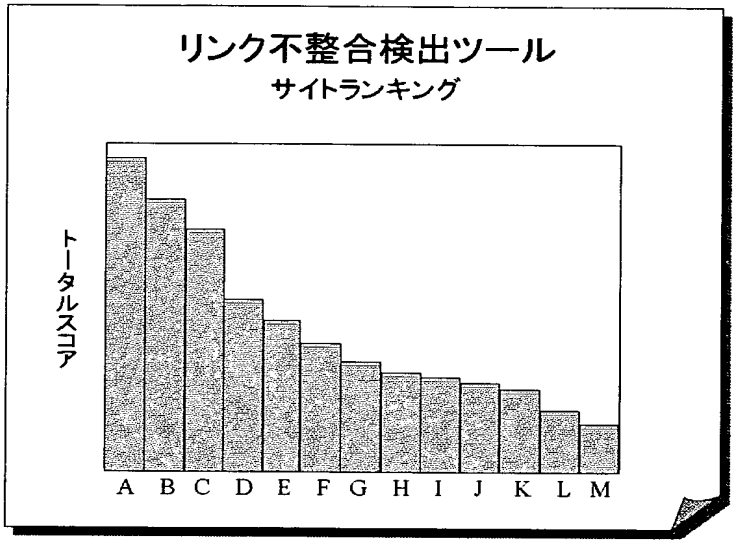
【図３０】

【図30】



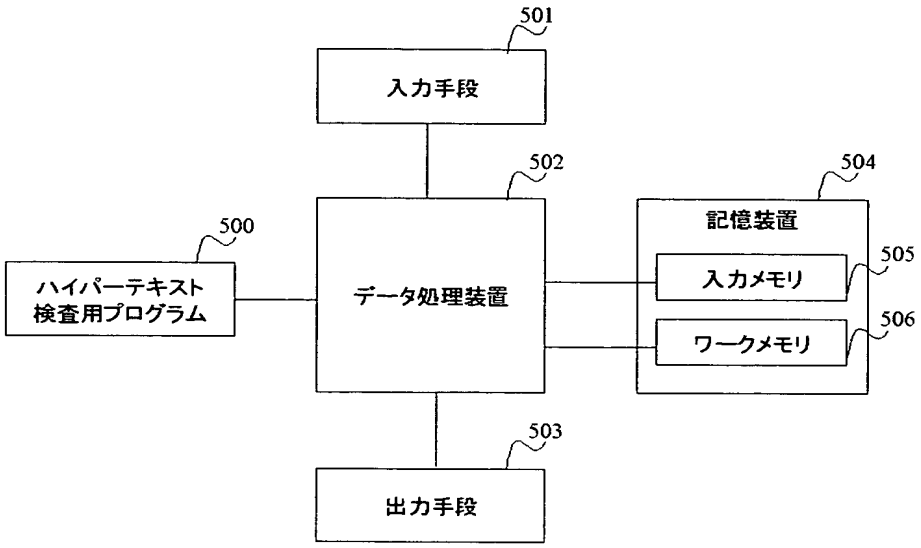
【図３１】

【図31】



【図３２】

【図32】


501
入力手段
502
504
記憶装置
500
ハイパーテキスト
検査用プログラム
データ処理装置
505
入力メモリ
506
ワークメモリ
503
出力手段

【書類名】　要約書

【要約】

【課題】　ハイパーテキストデータベースを対象とし、論理的に不整合なリンク箇所およびその訂正候補を自動的に求めて訂正する。

【解決手段】　情報収集手段１１はハイパーテキストを構成するページおよびリンクに関する情報をハイパーテキストデータベースベース２１から収集して情報記憶部２２に記憶する。条件判定手段１３は、情報記憶部２２を参照し、リンク情報を項目毎にグループ化し、グループから外れた特異なリンクをリンク不整合として抽出する。候補計算手段１２は、条件判定手段１３が抽出した特異なリンクのリンク情報を、グループと同一のリンク情報にするような訂正候補を計算する。訂正反映手段１４は、条件判定手段１３が検出したリンク不整合の箇所と候補計算手段１２が計算した訂正候補とに基づいてハイパーテキストデータベース２１を更新する。

【選択図】　図１

特願２００２－３０２５８５

出　願　人　履　歴　情　報

識別番号　　［０００００４２３７］

１．変更年月日　　１９９０年　８月２９日
［変更理由］　　新規登録
住　所　　東京都港区芝五丁目７番１号
氏　名　　日本電気株式会社